UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

MAYARA MACEDO DA MATA

ESPECTROSCOPIA RAMAN (PORTÁTIL) E QUIMIOMETRIA PARA DISTINÇÃO ENTRE UMA CULTIVAR DE ALGODÃO CONVENCIONAL E UMA GENETICAMENTE MODIFICADA

CAMPINA GRANDE
2021

MAYARA MACEDO DA MATA

ESPECTROSCOPIA RAMAN (PORTÁTIL) E QUIMIOMETRIA PARA DISTINÇÃO ENTRE UMA CULTIVAR DE ALGODÃO CONVENCIONAL E UMA GENETICAMENTE MODIFICADA

Trabalho de Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de mestre em Química.

Área de concentração: Química Analítica.

Orientador: Prof.ª Dr.ª Simone da Silva Simões.

Coorientador: Prof. Dr. Everaldo Paulo de Medeiros.

CAMPINA GRANDE
2021

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

M425e Mata, Mayara Macedo da.
Espectroscopia Raman (Portátil) e quimiometria para distinção entre uma cultivar de algodão convencional e uma geneticamente modificada [manuscrito] / Mayara Macedo da Mata. - 2021.
62 p.

Digitado.
Dissertação (Mestrado em Química - Mestrado) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2021.
"Orientação : Profa. Dra. Simone da Silva Simões , Departamento de Química - CCT."

1. Cultura do algodão. 2. Transgênicos. 3. Quimiometria. I. Título

21. ed. CDD 543

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

MAYARA MACEDO DA MATA

ESPECTROSCOPIA RAMAN (PORTÁTIL) E QUIMIOMETRIA PARA DISTINÇÃO ENTRE UMA CULTIVAR DE ALGODÃO CONVENCIONAL E UMA GENETICAMENTE MODIFICADA

Trabalho de Qualificação apresentada ao Programa de Pós-Graduação em Química da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de mestre em Química.

Área de concentração: Química Analítica.

Aprovada em: 15/03/2021.

BANCA EXAMINADORA

Prof. Dr.ª Simone da Silva Simões. (Orientadora)
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. José Germano Veras Neto – Examinador Interno
Universidade Estadual da Paraíba (UEPB/PPGQ)

Profª. Drª. Claudete Fernandes Pereira – Examinador Externo
Universidade Federal de Pernambuco (UFPE/DQF)

AGRADECIMENTOS

Agradeço a Deus por ser essencial e luz na minha vida, por toda trajetória conquistada até aqui, por ter me dado força e coragem para enfrentar as diversidades da vida e por me dar clareza de pensamentos para aprender dia após dia.

A minha mãe Marquesa e a minha irmã Maynara, por me motivarem sempre em busca dos meus sonhos. Amo muito vocês.

A minha família pelo apoio, incentivo e por fazerem parte desta história de conquistas.

A meu namorado Wylliam Eduardo por todo companheirismo, afeto e amor. Obrigada por aguentar minhas crises de estresse e ansiedade e por estar presente nos momentos mais importantes para mim. Amo você!

As minhas amigas Samara, Louyse, Adna e Alline pelo seu carinho, estímulo, amizade, companheirismo e pelas palavras de motivação que me fizeram ter forças para seguir em frente.

As amizades que o mestrado me proporcionou e que foram enviadas como presente de Deus: Lucilane, Diego, Luana, Welida, Jéssica e Odilon. Vocês fizeram essa trajetória ser mais leve. Obrigada por não soltarem a minha mão.

A Profª. Drª Simone da Silva Simões por toda orientação e pelos ensinamentos que irei levar para vida profissional e pessoal e que foram fundamentais para a realização deste sonho. Obrigada por todo apoio.

A todos os professores e funcionários da secretaria do PPGQ que tiveram sua cota de contribuição na finalização nesta etapa da minha vida.

Ao Núcleo de Química Analítica Avançada do Estado de Pernambuco por essa parceria para a construção e andamento desse trabalho.

E agradeço a todos que de forma direta ou indiretamente contribuíram e fizeram parte deste ciclo que se encerra, o meu muito obrigado.

“Algumas pessoas marcam a nossa vida para sempre, umas porque nos vão ajudando na construção, outras porque nos apresentam projetos de sonho e outras ainda porque nos desafiam a construí-los”.

“Mentalize um futuro para sua vida e não pare até chegar lá”. (Rafael Magalhães)

RESUMO

Os transgênicos passam por alterações genéticas de modo a explorar características agronômicas superiores por meio de métodos de engenharia genética. Técnicas biotecnológicas são utilizadas para o melhoramento de espécies como também, resolver problemas relacionados a qualidade, produtividade e resistência a pragas. O algodão como uma das principais comodities para o país detêm um alto nível tecnológico que garante ser produzido com sustentabilidade ambiental (menor pegada hídrica do mundo), social e de qualidade de fibras. Ao mesmo tempo, o Brasil já se destaca com grande potencial para o seguimento de algodão orgânico em áreas do semiárido, envolvendo pequenos produtores. Como fundamentação, os novos padrões de consumo da sociedade exigem garantias sustentáveis de produção e isso requer como condição essencial a utilização de métodos analíticos exploratórios, rápidos e economicamente viáveis. A espectroscopia vibracional tem um conjunto de técnicas com características atrativas para o desenvolvimento de métodos para atender a esses requisitos tecnológicos e comerciais. Dentre elas, a espectroscopia Raman como instrumentação analítica e explorada utilizando modelos multivariados é uma alternativa que possui vantagens para estudos de implementação. O objetivo deste trabalho foi desenvolver métodos para distinção de sementes em cultivares de algodão convencional e transgênicas utilizando espectroscopia Raman e técnicas de reconhecimento de padrões. Foram utilizadas 20 amostras de sementes individuais de algodão (10 convencionais e 10 transgênicas). Os espectros Raman de todas as amostras foram registrados e um modelo PLS-DA foi construído para distinção entre as sementes de algodão. Na etapa de treinamento foi obtido 97,1% de acertos e um RMSE de 0,029 e na etapa de teste foi obtido 96,7% de acertos e um RMSE de 0,033. A distinção entre cultivares convencionais e transgênicas de algodão por meio de modelos de reconhecimento de padrão juntamente com a espectroscopia Raman portátil, é uma alternativa como método de diagnóstico rápido, fácil implementação e econômico que permite ser explorado e validado para implementação em larga escala.

Palavras-Chave: Cultura algodoeira. Transgênicos. Quimiometria.

ABSTRACT

Transgenics undergo genetic changes in order to exploit superior agronomic traits through genetic engineering methods. Biotechnological techniques are used for the improvement of species as well as to solve problems related to quality, productivity and resistance to pests. Cotton as one of the main commodities for the country has a high technological level that guarantees it is produced with environmental sustainability (the smallest water footprint in the world), social sustainability, and fiber quality. At the same time, Brazil already stands out with great potential for organic cotton in semi-arid areas, involving small producers. As a rationale, society's new consumption patterns require sustainable production guarantees, and this requires as an essential condition the use of exploratory, rapid, and economically feasible analytical methods. Vibrational spectroscopy has a set of techniques with attractive characteristics for developing methods to meet these technological and commercial requirements. Among them, Raman spectroscopy as analytical instrumentation and explored using multivariate models is an alternative that has advantages for implementation studies. The objective of this work was to develop methods for distinguishing seeds in conventional and transgenic cotton cultivars using Raman spectroscopy and pattern recognition techniques. Twenty individual cotton seed samples (10 conventional and 10 transgenic) were used. Raman spectra of all samples were recorded and a PLS-DA model was constructed to distinguish between cottonseeds. In the training stage 97.1% of hits and an RMSE of 0.029 were obtained, and in the test stage 96.7% of hits and an RMSE of 0.033 were obtained. Distinguishing between conventional and transgenic cotton cultivars by means of pattern recognition models together with portable Raman spectroscopy is an alternative as a fast, easily implemented and cost-effective diagnostic method that can be explored and validated for large-scale implementation.

Keywords: Cotton culture. Transgenics. Chemometrics.

LISTA DE ILUSTRAÇÕES

Figura 1 – Diagrama do nível de energia do espalhamento Raman..... 18

Figura 2 – Organização dos dados para calibração multivariada..... 28

Figura 3 – Curva ROC hipotética 33

Figura 4 – Espectrômetro Raman Portátil..... 35

Figura 5 - Sementes de algodão convencionais e transgênicas..... 36

Figura 6 - Espectros brutos das sementes convencionais e transgênicas..... 36

Figura 7- Espectros brutos sem as regiões externas das sementes convencionais e transgênicas..... 37

Figura 8 - Espectros brutos com as 120 amostras (a) e espectros removidos do banco de dados (b)..... 38

Figura 9 - Dados pré-processados com a) NVU+WLS; b) NVU+WLS +suavização com polinômio Savitzky-Golay de 1ª ordem e janela de 15 pontos; c) NVU+WLS +suavização com polinômio Savitzky-Golay de 1ª ordem e janela de 51; d) NVU+WLS +suavização com polinômio Savitzky-Golay de 1ª ordem e janela de 15 pontos + SNV ; e) NVU+WLS +1ª derivada com polinômio Savitzky-Golay de 1ª ordem e janela de 15 pontos ; e f) NVU+WLS +1ª derivada com polinômio Savitzky-Golay de 2ª ordem e janela de 51 pontos..... 39

Figura 10 - Gráfico dos escores para a predição das amostras tratadas com a) NVU+WLS, b) NVU+WLS +suavização com polinômio SG de 1ª ordem e janela de 15 pontos, c) NVU+WLS +suavização com polinômio SG de 1ª ordem e janela de 51, d) NVU+WLS +suavização com polinômio SG de 1ª ordem e janela de 15 pontos + SNV utilizando apenas a região destacada em anterior (Figura 9d); e) NVU+WLS +1ª derivada com polinômio SG de 1ª ordem e janela de 15 pontos; e f) NVU+WLS +2ª derivada com polinômio SG de 1ª ordem e janela de 51..... 40

Figura 11 - T^2 Hotteling versus Q Resudual..... 43

Figura 12 - Números de Variáveis Latentes..... 43

Figura 13 - Predição das classes 44

Figura 14 - Curva ROC do modelo PLS-DA..... 45

Figura 15 - a) Espectros pré-processados, b) até f) loadings nas variáveis latentes de 1 a 5, respectivamente e g) escores VIP´s. (I), (II) e (III) se referem aos comprimentos de onda 1330 cm^{-1}, 1342 cm^{-1} e 1656 cm^{-1}, respectivamente, destacado com a linha hachurada em vermelho. As faixas em azul destacam as regiões importantes de acordo com o gráfico de VIPs escores..... 46

Tabela 1 – Figuras de Mérito de cada pré-processamento. 42

Tabela 2 – Cronograma das atividades a serem desenvolvidas durante a vigência do projeto Mar/2019 a Fev/2021. 60

LISTA DE ABREVIATURA

ALS – Mínimos quadrados assimétricos (*Asymmetric Least Squares*)

ANVISA - Agência Nacional de Vigilância Sanitária

CLS- Mínimos quadrados clássicos (*classical least squares*)

CTNBio – Comissão Técnica Nacional de Biossegurança

DNA – Ácido Desoxirribonucleico (*Deoxyribonucleic Acid*)

ELISA – Ensaio de Imunoabsorção Enzimática (*Enzyme-Linked Immunosorbent Assay*)

EMBRAPA- Empresa Brasileira de Pesquisa Agropecuária

FN – Falso Negativo

FP- Falso Positivo

FT-RAMAN - Espectroscopia Raman por Transformada de Fourier

HCA – Análise Hierárquica de Cluster

IFL – Imunoensaio de Fluxo Lateral

INMETRO - Instituto Nacional de Metrologia, Qualidade e Tecnologia

IV – Infravermelho

KNN – K Vizinhos mais próximos *(k-nearest neighbor)*

LDA – Análise Discriminante Linear

LOD – Limite de Detecção

MLR – Regressão Linear Múltipla

MSC- Correção de sinal multiplicativa *(Multiplicative Signal Correction)*

NAS- Sinal Analítico Líquido (*Net Analyte Signal*)

NIR – Infravermelho Próximo (*Near Infrared*)

NVU – Normalização pelo Vetor Unitário

OD- Distância Ortogonal

OGM's – Organismos Geneticamente Modificados

PCA – Análise de Componentes Principais (*Principal Component Analysis*)

PCR - Reação em Cadeia da Polimerase (*Polymerase Chain Reaction*)

PLS – Mínimos Quadrados Parciais (*Patial Least Squares*)

PLS-DA – Análise Discriminante por Mínimos Quadrados Parciais (*Partial Least Squares – Discriminant Analysis*)

RMSEC- – Raiz Quadrada do Erro Médio Quadrático da Calibração *(Root Mean Square Error of Calibration)*

RMSEP – Raiz Quadrada do Erro Médio Quadrático de Previsão *(Root Mean Squares Error of Prediction)*

SD – Distância Scores

SIMCA – Modelagem Independente Flexível por Analogia de Classes *(Soft Independent Modeling of* Class *Analogy)*

SNV- Transformada padrão normal (*Standard Normal Variate Transformation*)

VL – Variáveis Latentes (*Latent Variables*)

VN – Verdadeiro Negativo

VP – Verdadeiro Positivo

WLS – Mínimos quadrados ponderados *(Weighted Least Squares*)

SUMÁRIO

1 INTRODUÇÃO 12

2 OBJETIVOS..... 14

2.1 Objetivo geral 14

2.2 Objetivo específico 14

3 REVISÃO DA LITERATURA 15

3.1 Algodão..... 15

3.2 Espectroscopia Raman..... 16

3.3 Uso da Espectroscopia vibracional para monitoramento de características genéticas em produtos agrícolas 19

3.4 Tratamento quimiométricos de dados obtidos por Espectroscopia Raman..... 20

3.4.1 Construção de Modelos Multivariados..... 27

3.4.1.1 Regressão por Mínimos Quadrados Parciais (PLS)..... 28

3.4.1.2 Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA)..... 29

3.5 Figuras de Mérito..... 30

4 MATERIAL E MÉTODOS 34

4.1 Amostras..... 34

4.2 Instrumentação Analítica..... 34

4.3 Tratamento de dados..... 35

5 RESULTADOS E DISCUSÃO..... 36

5.1 Análise dos espectros Raman..... 36

5.2 Pré-processamento dos dados Raman..... 37

5.3 Modelo PLS-DA..... 40

6 CONCLUSÃO 48

REFERENCIAS 49

ANEXO A- CRONOGRAMA DE ATIVIDADES DO TRABALHO DE DISSERTAÇÃO..... 60

ANEXO B- CURRICULUM LATTES..... 61

1 INTRODUÇÃO

O algodão é uma das culturas mais relevante na agricultura de vários países, com impacto econômico para a balança comercial de países em desenvolvimento (AHMAD *et al*., 2017; AHMAD e RAZA, 2014; ABBAS e AHMAD, 2018). O algodão é considerado a melhor fibra natural comercial produzida em larga escala e a principal fonte de fibra fiáveis para produtos têxteis (PETERSON *et al.,* 2012).

A fibra protege a semente de algodão contra a entrada de insetos, fazendo uma barreira protetora para apoiar a germinação e propagação da semente em condições ambientais favoráveis (CANILHA *et al.,* 2010; SATYAMURTHY *et al.,* 2011; MEYABADI *et al.,* 2014). E a qualidade das fibras depende de fatores bióticos e abióticos que poderá influenciar a sua industrialização e comercialização (FREDDI; STOQUERO; PERES, 2014).

O desenvolvimento e crescimento do algodoeiro depende da configuração da semeadura que provoca mudanças nas características fisiológicas e morfológicas da cultura e da planta (FLOWLER; RAY, 1977). Uma problemática que envolve à cultura de algodão, seja no semiárido no bioma da caatinga ou cerrado brasileiro, com mais intensidade, é a presença frequente de pragas e doenças com capacidade de gerar danos, como: o bicudo, pulgão, lagarta, mosca-rosa, a manha de ramulária e outros, que a cada safra causam altos custos e riscos eminentes à cultura (SANTOS, 2007).

Através da engenharia genética diferentes variedades de algodão vêm sendo desenvolvidas, o que resulta no aumento de opções tecnológicas aos sistemas de produção de algodão. A modificação genética é uma biotecnologia importante que é explorada em diversas aplicações, a qual permite para conferir resistência ou tolerância a herbicidas, com eventos múltiplos já envolvendo insetos e pragas na agricultura (LI *et al.,* 2014). Segundo Ahmad et. al (2017) com o avanço das técnicas de melhoramento genético no meio agrícola e com o surgimento de cultivares tolerantes a herbicidas e a insetos-pragas, essas técnicas vêm constituindo uma importante ferramenta para o manuseio de algumas culturas. Essas técnicas trazem como benefício a redução do uso de produtos fitossanitários, perdas e custos com manuseio, bem como maior segurança de cultivo.

Embora haja evidências dos benefícios sobre as culturas agrícolas com eventos transgênicos, seu crescimento em larga escala e a perda das características tecnológicas incorporadas ao longo tem suscitado estudos envolvendo a rastreabilidade e classificação

(LU e YANG, 2009; LU et al., 2010). Nesse sentido, a detecção e identificação de materiais com tais eventos de outros não incorporados é um dos requisitos fundamentais para a segurança de tais tecnologias, assim como da sociedade beneficiada.

Como alternativa aos métodos tradicionalmente empregados na detecção e rastreabilidade de produtos transgênicos as tecnologias consideradas não destrutivas têm demonstrado resultados na identificação rápida que incluem tecnologias de espectroscopia de infravermelho próximo, visível e imagens multiespectrais combinado técnicas quimiometricas (XIE *et al.*, 2007; LIU *et al.*, 2017, ROCHA, 2020).

Diante do exposto, a espectroscopia Raman pode ser vista como uma ferramenta alternativa aos métodos convencionais de análise. Deste modo, o presente trabalho propõe o desenvolvimento de métodos baseados na espectroscopia Raman e quimiometria para identificação e distinção entre sementes convencionais e transgênicas de algodão.

2 OBJETIVOS

2.1 Objetivo geral

Aplicar a espectroscopia Raman e quimiometria para a distinção de sementes individuais de cultivares de algodão convencional e transgênico.

2.2 Objetivos específicos

- ✓ Avaliar o desempenho do espectrômetro RAMAN portátil na distinção entre sementes de algodão convencional e geneticamente modificadas;
- ✓ Avaliar o desempenho de diversos pré-processamentos espectrais para a remoção de características espectrais indesejadas e, consequentemente, melhoria dos modelos de reconhecimento de padrões;
- ✓ Construir e validar modelos de reconhecimento de padrões supervisionados utilizando espectros Raman no infravermelho próximo.

3 REVISÃO DA LITERATURA

3.1 Algodão

O algodão é uma espécie vegetal pertencente à família Malvaceae, que corresponde a 85 gêneros *Gossypium* e mais de 1.500 espécies (BAGATIN, 2008). Dentre outras espécies cultivadas destacam-se *Gossypium herbaceum* e *Gossypium arboreum* (PENNA, 2005; ZHANG; HENSON 2007). A cultura do algodão é uma planta com o sistema reprodutivo considerado misto, uma vez que ocorre simultaneamente a polinização entomófila e a autofecundação (CRISÓSTOMO, 1989).

O algodão é considerado como a principal cultura de fibras, sendo a mais importante fibra natural (PUJER *et al.*, 2014). A fibra protege a semente de algodão contra a entrada de insetos, fazendo uma barreira protetora para apoiar a germinação e propagação da semente em condições ambientais favoráveis (CANILHA *et al.*, 2010; SATYAMURTHY *et al.*, 2014; MEYABADI *et al.*, 2014).

Diferentes cultivares de algodão vêm sendo estudadas para garantir uma melhor adaptabilidade as adversidades climáticas, resistência ao ataque de doenças e pragas, bem como verificar o rendimento em plumas e o crescimento e produção de capulhos. Esse tema é o foco principal das pesquisas desenvolvidas por diversos centros de pesquisa como a Embrapa Algodão (CARVALHO; ANDRASE; SILVA FILHO, 2011).

A cultura algodoeira em condições tropicais atrai variedades de pragas e doenças (fungos e bactérias) que exigem um controle intensivo, sobretudo, com fitossanitários que podem afetar o meio ambiente e insetos benéficos. Deste modo, faz-se uso de produtos químicos seletivos para o controle de insetos e doenças à cultura, de forma que ofereça controle de pragas, ao mesmo tempo menor impacto ambiental (MACHADO et al., 2019).

Portanto, é necessário que as estratégias de manejo de pragas sejam bem executadas para não acontecer perdas no rendimento (FAND; SUROSHE, 2015; KAMBLE; FAND, 2019).

O algodão Bt (*Bacillus thuringiensis*) tem sido relevante, proporcionando aumento da produtividade do algodão, limitando as principais pragas, diminuindo pulverizações de inseticidas e melhorando o controle de pragas por inimigos naturais (KRANTHI; RUSSEL, 2009).

Plantas transgênicas contendo genes do *Bacillus thuringiensis* (Bt) que produzem toxinas inseticidas e têm sido usadas em programas de manejo de pragas e sua principal vantagem econômica é a redução dos custos com inseticidas, bem como melhores

vantagens ambientais (BORGES, 2010). Por meio de técnicas biotecnológicas os genes da toxina são inseridos em uma variedade de culturas. Os cultivos Bt expressam o gene inseticida em todo o seu ciclo de crescimento. Como exemplo o algodão Bt é eficiente para combater pragas de lagartas que prejudicam a produção em diversas regiões produtoras (SILVEIRA, 2001).

Outros exemplos de materiais transgênicos como a cultivar transgênica BRS 368 RF a qual é resistente ao herbicida glifosato, permite maior flexibilidade no controle de plantas daninhas dispensando a utilização de pulverizações com herbicidas não seletivos, sendo possível a aplicação em qualquer fase do desenvolvimento da cultura. Essa cultivar apresenta características como: o ciclo de médio a precoce, porte baixo, rendimento de fibra de 40% e potencial produtivo entre 4.200 a 4.500 Kg/ha (EMBRAPA, 2013). Destaca-se por seu alto rendimento de fibra e atributos de qualidade para a fibra (BARROSO; FREIRE; AMARAL., 2005).

De maneira geral, a biotecnologia tem adquirido espaço na agricultura e a engenharia genética é um dos temas que mais se destacam, principalmente devido às aplicações nos setores agrícolas. O desenvolvimento de transgênicos tem sido utilizada especialmente no desenvolvimento de culturas economicamente importantes, tais como o algodão. O desenvolvimento da engenharia genética e o uso das culturas transgênicas têm como objetivo obter plantas resistentes a pragas e doenças, sendo necessária uma menor aplicação de agrotóxicos durante a produção agrícola (QAIM; ZILBERMAN,2003). Segundo Borges (2010) os cultivos transgênicos vêm adquirindo aceitação por partes dos agricultores e essa aceitação está relacionada aos benefícios financeiros e não financeiros que as sementes geneticamente modificadas em relação as sementes transgênicas proporcionam.

3.2 Espectroscopia Raman

As técnicas espectroscópicas são baseadas na capacidade dos átomos e moléculas poderem emitir, absorver e espalhar energia em determinadas regiões dos espectros eletromagnéticos, proporcionando informações sobre a composição química e cristalina da amostra (LI-CHAN *et al.*, 1994; FERRARO, 2003). Essas técnicas são utilizadas na caracterização e identificação de estruturas de compostos e moléculas (HERRENO, 2008). Essas técnicas são ferramentas eficazes na caracterização de materiais em diversas áreas do conhecimento (FARIA *et al.*, 2002). Algumas técnicas espectroscópicas são não destrutivas e fornecem informações da amostra, com mínimo ou até mesmo nenhum

preparo da amostra em um curto período de tempo (ABBAS; DARDENNE; BAETEN, 2012).

A espectroscopia vibracional fornece uma grande quantidade de detalhes sobre o material analisado, medindo os níveis de energia vibracionais através da radiação eletromagnética, associados às ligações químicas na amostra (SALA, 2008; FARIA; SANTOS; GONÇALVES, 1997). A espectroscopia vibracional é composta pelas técnicas no infravermelho, Raman e Terahertz. A intensidade dos espectros do infravermelho e Raman será diferente devido o fenômeno físico que comanda cada uma delas. Assim, essas duas espectroscopias se tornam complementares na investigação das vibrações moleculares, sendo frequentemente aplicadas na consecução de informações sobre a composição química e estrutura molecular (FERRARO , 2003; SALA, 2008).

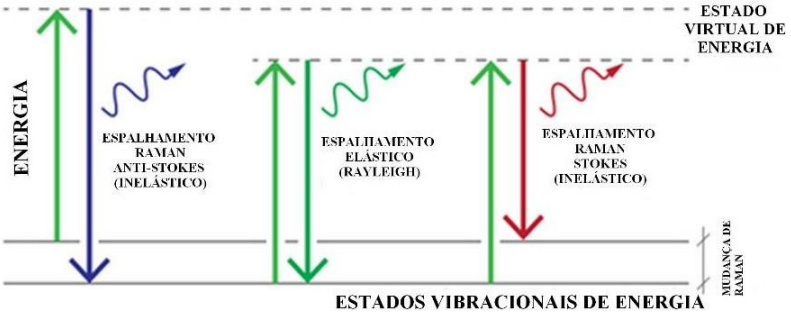
A espectroscopia Raman caracteriza substâncias com base no espectro obtido a partir dos feixes espalhados que pode ser associado com à composição química e estrutura molecular das amostras (SANTOS *et al.,* 2019). Foi descoberta em 1928 pelo físico indiano Chandrasekhara Venkata Raman quando estudava o espalhamento da luz, argumentando sobre a natureza quântica da luz no experimento de Compton. Dada a reconhecida importância da descoberta, foi-lhe concebido o Prêmio Nobel de Física em 1930 (ULF, 1993).

Baseias-se nas transições vibracionais discretas que acontecem no estado eletrônico da molécula, fornecendo informações sobre seus modos vibracionais. Dessa forma, quando a amostra é irradiada com uma luz monocromática visível ou infravermelha próxima de um laser, é possível obter algumas informações físicas do composto determinado, como também, informações sobre a estrutura molecular a partir da radiação difratada (ou espalhada). Isso acontece devido os níveis de energia vibracional na molécula passar de um estado de colisão de curta duração e alta energia, retornando a um estado de energia mais baixo pela emissão de um fóton (HERRENO, 2008).

A dispersão Raman compreende uma ampla variedade de fenômenos de dispersão da luz, ocorrendo alterações no comprimento de onda quando as partículas dispersantes forem pequenas comparadas com o comprimento de onda da luz, sendo exposto como efeito de dispersão inelástica da luz (ULF, 1993). A natureza da luz dispersa depende do tipo de amostra presente, reforçando a importância desse fenômeno como uma ferramenta de pesquisa (BOHNING; MISHRA; CHOUDHURY, 1998).

O espalhamento Raman pode ocorrer na forma elástica (Rayleigh) e inelástica. No espalhamento Rayleigh (energia elástica) o feixe incidente e a radiação espalhada possuem a mesma frequência, nele uma determinada amostra recebe radiações eletromagnéticas com uma determinada frequência, sendo elevada a um estado vibracional virtual e retornando para o mesmo estado vibracional inicial. Isso implica na emissão de uma radiação eletromagnética de mesma frequência e mesma energia que a incidida, conforme Figura 1. O espalhamento Stokes ocorre quando a radiação espalhada é de frequência inferior que a incidente, ou seja, quando uma molécula é excitada do seu estado fundamental para um estado virtual, podendo absorver ou emitir um fóton de energia uma pequena parte da radiação é espalhada inelasticamente. A molécula também pode estar em um nível vibracional mais elevado, diante disto a radiação espalhada com frequência maior que a fonte de excitação é chamada de espalhamento anti-stokes (HOLLER; SGOOK; CROUCH,2009).

Figura 1 - Diagrama do nível de energia do espelhamento Raman



Fonte: Adaptado de Jablonski (1933).

O espalhamento anti-stokes apresenta picos menos intensos que os picos correspondentes ao espalhamento Stokes, por este motivo os equipamentos Raman consideram somente a região Stokes do espectro. Os espectros Raman são geralmente apresentados com a abcissa definida pelo deslocamento do número de onda (cm^{-1}) que é proporcional a diferença de energia entre a radiação espalhada e a da fonte, a ordenada é representada pela intensidade do sinal (SGOOK; HOLLER; CROUCH,2009). Sendo assim, a espectroscopia Raman, geralmente, apresenta uma natureza não destrutivacomo uma de suas principais vantagens, como também considerada uma técnica rápida. Entretanto, uma limitação para esta técnica é a baixa qualidade de sinal para amostras opacas ou escura, bem como a presença de ruídos ocasionados pela fluorescência (SANTOS *et al.*, 2019).

A espectroscopia Raman registra a intensidade da radiação espalhada pelas moléculas, resultando na alteração da polarização das ligações devido ao campo elétrico da radiação incidente, ocasionando a um estado eletrônico virtual. A intensidade Raman depende da polarizabilidade da molécula, concentração do grupo molecular, intensidade da fonte e isso torna-o bastante complexo. As intensidades das bandas Raman são diretamente proporcionais à concentração do analito na amostra. Para uma mesmaamostra os espectros Raman apresentam menos bandas que os espectros no infravermelho, tornando mais simples as medidas quantitativas (BOUGEARD; SCHADER, 1995)

3.3 Uso da espectroscopia vibracional para monitoramento de características genéticas em produtos agrícolas

As técnicas espectroscópicas vibracionais são ferramentas de detecção não destrutivas e altamente utilizadas devido a fácil utilização, rapidez e simplicidade. De maneira geral, as técnicas espectroscópicas permitem investigar uma amostra para obter informações de caráter qualitativo ou quantitativo a partir de informações detalhadas sobre os níveis de energia das espécies em estudo. A espectroscopia vibracional proporciona maior riqueza de detalhes, devido as transições vibracionais e rotacionais, obtendo orientações sobre a natureza das ligações químicas e geometria molecular, tais informações são caracterizadas através dos espectros (FARIA; SANTOS; GONÇALVES, 1997). A espectroscopia Raman é um tipo de espectroscopia vibracional que emprega a interação da radiação eletromagnética com a matéria, obtendo informações sobre a estrutura molecular e as ligações químicas presentes (SILVA et. al, 2013). A detecção de material transgênico através do uso das espectroscopias no infravermelho próximo, visível e Raman vem sendo comprovadas por algumas publicações e destacam o êxito dessas ferramentas (LEVANDI *et al.*, 2008; XU *et al.*, 2014). Essas técnicas são normalmente utilizadas aliadas à métodos de análise multivariada para análise direta das amostras (LAVINE e WORKMAN, 2013). Neste sentido, diversos estudos têm investigado os organismos transgênicos a partir de técnicas espectroscópicas.

Xu et al.(2014) utilizaram técnicas espectroscópicas e análise multivariada para identificar organismos transgênicos, dispondo dos métodos quimiometricos para classificação na análise da obtenção dos espectros para diferenciação de arroz

transgênico. Assim, os autores observaram que o modelo de análise discriminante de mínimos quadrados parciais e a espectroscopia Raman apresentaram resultados com uma maior precisão, mostrando que a espectroscopia Raman, juntamente com os métodos quimiometricos podem ser eficazes na diferenciação do arroz transgênico.

Baranski e Baranska, (2008) aplicaram a espectroscopia FT-Raman para discriminar os tecidos transgênicos e os materiais que não pertenciam a essa classe, mostrando os rearranjos genéticos. Através da análise de agrupamentos analisaram tecido vegetal de cenouras transgênicas e de folhas de sementes de tabaco, permitindo obter uma separação das amostras, mostrando uma discriminação entre as espécies transgênicas e os materiais que não pertenciam a essa classe. Os autores concluíram que a espectroscopia Raman apresenta é uma técnica valiosa para a discriminação de sementes transgênicas, obtendo informações sobre a condição genético sem a necessidade de destruição da amostra (BARANSKI; BARANSKA, 2008).

Liu (2016) utilizou a quimiometria acoplada a espectroscopia Terahertz para identificar óleo de milho transgênico, mostrando que os modelos de análise discriminante por mínimo quadrados parciais (PLS-DA) aplicados a dados espectroscópicos conseguem classificar óleos transgênicos e não transgênicos com precisão, porém sendo melhor para o conjunto de validação. Nessa mesma perspectiva, Xu et al. (2015) usou a espectroscopia Terahertz combinadas com métodos quimiometricos como ferramentas para discriminar propriedades semelhantes da proteína transgênica. Os resultados mostraram que as técnicas espectroscópicas e os métodos quimiometricos podem ser novas formas possíveis de diferenciar essas proteínas. As técnicas espectroscópicas oferecem benefício de evitar trabalhos demorados e análises químicas e sensoriais caras. O uso dessa técnica para determinar amostras transgênicas é benéfico para esse ramo, mostrando potencial aplicação para a discriminação de produtos transgênicos.

3.4 Tratamento quimiometrico de dados obtidos por espectroscopia Raman

A quimiometria utiliza métodos matemáticos e estatísticos com o objetivo de planejar ou otimizar procedimentos experimentais para extrair o máximo de informações de dados químicos, possibilitando a análise de dados de maior complexidade. Essa área é direcionada aos estudos de dados químicos e o seu surgimento veio com a evolução de equipamentos de análise de dados. Tais equipamento são fundamentados na matemática, estatística e computação, mas os problemas que se visa resolver são de origens e de interesse química (ALMEIDA,et. al, 2015; BOVOLINI, 2010).

Os dados Raman são baseados no efeito Raman, isto é, a radiação eletromagnética sofre espalhamento inelástico. Esse espalhamento possui intensidade relativamente baixa e polarização. Deste modo, o tratamento dos dados obtidos requer etapas de pré-processamento diferentes daquelas utilizadas em espectroscopia no infravermelho (IV) (BURGER; GELAD, 2005). Além disto, os espectros Raman são afetados pela interferência de raios cósmicos que são portadores de cargas que ocasionam discrepância no espectro, geralmente aparecem em picos estreitos, positivos e com largura de banda e em posições aleatórias (ZHANG e HENSON, 2007).

A estatística multivariada é indispensável no tratamento de dados espectrais Raman, pois esses métodos permitem a construção de modelos capazes de distinguir diferenças espectrais de amostras (HORTOR *et al.*, 2011; O'CONNELL *et al.*, 2010). A correção de linha de base pode ser feita pela aplicação de polinômios como mínimos quadrados ponderados ou assimétricos (EIERS E BOELENS, 2005) e por métodos de derivação (SMITH; HUFF; GORDON, 2016). Os problemas envolvendo diferenças de escala podem ser corrigidos pelos métodos de normalização acentuando o vetor unitário. Os problemas que acarretam mudança na linha de base são ocasionados pela fluorescência e dispersão que induzem a mudança desiguais de amplitude em diferentes números de ondas e normalmente não estão relacionados a composição química da amostra. Os valores calculados da mediana envolvem uma função gaussiana para evitar a descontinuidade da linha de base (VIDAL; AMIGO, 2012; FRIEDRICHS, 1995).

A seleção incorreta da ordem polinomial pode afetar algumas bandas Raman, como também afetar ruídos de alta frequência e, consequentemente, poder apresentar dados incoerentes (ZHAO *et al.*, 2007; ESMONDE-WHITE; ESMODE-WHITE; MORRIS, 2011). Além disso, o registro dos espectros de um determinado material pode ser realizado em momentos e condições distintas, produzindo disparidade nos níveis de intensidade. Com isso, o processo de normalização é uma etapa importante do pré-processamento, sendo responsável pelo tratamento dessa diferença nos níveis de intensidade. Sendo assim, esse tipo de processamento exerce a função de verificar a intensidade de uma determinada banda Raman entre os espectros registrados é o mais semelhante possível (BOCKLITZ *et al.*, 2011; RANDOLPH, 2006).

A presença de picos cósmicos dificulta a análise dos dados multivariados, podendo inserir variáveis equivocadas nas técnicas de regressão e nas curvas multivariadas, bem como, podendo ocasionar erros na análise de classificação e causando distorções nas direções da análise de componentes principais (PCA) (ZHANG e

HENSON, 2007). Os raios cósmicos (*spikes*) são picos estreitos de intensidade variada e geralmente possuem semelhança às bandas Raman que causam interferências nos espectros. A identificação e eliminação desses raios (*spikes*) presentes nos espectros são fundamentais, visto que podem camuflar ou corromper bandas de interesse químico, limitando a execução dos métodos de análise de dados.

Algoritmos de suavização ou aplicação de filtros medianos podem auxiliar na remoção ou detecção dos raios cósmicos. Porém, como a resolução espectral Raman apresenta um nível alto, levando a espectros estreitos, os algoritmos medianos podem interferir nesses tipos de bandas. Assim, métodos de filtragem baseado na resolução espectral, podem ser usados para filtrar todos as características intensas que são mais estreitas que a resolução espectral inerente (KATSUMOTO e OZAKI, 2003**;** LAGO; JULIANO; KASCHERES, 1995; ZHANG e BEN-AMOTZ, 2002).

O sinal bruto do Raman possui ruídos que expandem componentes de alta e baixa frequência e a diferenciação amplifica essas componentes. A linha de base também contém componentes de baixa frequência comparados às bandas Raman. O ruído nos dados pode ser corrigido pela execução de métodos de suavização Savitzky-Golay e suavização por Kernel que são aplicados após a diferenciação e são baseados na aplicabilidade de um filtro polinomial. Relativamente, polinômios de menor grau e janelas maiores oferecem uma maior suavização, porém, a distorção desses parâmetros pode ocasionar desaparecimento de picos importantes (GAUTAM *et al.*, 2015; SACRÉ *et al.*, 2014)**.**

A etapa de pré-processamento é fundamental para o sucesso da análise multivariada, fazendo com que as condições da matriz de dados sejam melhoradas para que as informações sejam exploradas com eficiência. Métodos de processamento de dados como derivação e suavização usando algoritmo Savitzky-Golay, normalização dos espectros, correção de sinal multiplicativo e centralização dos dados na média são alguns métodos que podem ser utilizados para esse procedimento (RINNAN; VANDENBERG; ENGELSEN, 2009).

Os métodos estatísticos consideram amostras e variáveis, extraindo informações complementares dos dados, irrealizável pela análise univariada. A análise de componentes principais (PCA) e a análise hierárquica de clusters (HCA) pertencem a classe dos métodos não supervisionados em que, a separação das classes acontece sem a necessidade de informações sobre as amostras para a formação de agrupamentos entre as mesmas. Já a análise discriminante linear (LDA), algoritmo K-Nearest Neighbor (KNN),

análise discriminante por mínimos quadrados parciais (PLS-DA) e a modelagem independente suave por analogia de classe (SIMCA), pertencem aos métodos supervisionados, nos quais é necessário há existência de informações prévias sobre características das amostras para a construção do modelo (BRERETON, 2012).

O pré-processamento é uma manipulação matemática e é realizada após os dados serem organizados em matrizes. Geralmente, a matriz de dados precisa passar por um pré-processamento para remover ou reduzir efeitos irrelevantes para o objetivo da modelagem (BEEBE; PELL; SEASHOLTZ, 1998). Os pré-processamentos podem ser usados nas linhas e/ou colunas, isto é, nas variáveis e/ou nas amostras. Assim, quando a aplicação do pré-processamento acontece nas colunas, considera-se todas as amostras em uma variável por vez. Por sua vez, quando a aplicação acontece nas linhas, considera todas as variáveis em uma amostra por vez (FERREIRA, 2015). Dentre algumas etapas de pré-processamento, temos:

Normalização

As respostas analíticas obtidas pelos métodos instrumentais ficam dependentes de variações multiplicativas e aditivas, isto significa uma diminuição ou aumento no sinal medido. As variações de intensidade que acontecem por meio das fontes de radiação, mudanças experimentais ou pela resposta do detector são tipos de variações indesejadas. Essa variação não desejada em um determinado conjunto de dados pode ser solucionada por meio da aplicação de algum pré-processamento de normalização (SOUZA; POPPI, 2012; SHIPP; SINJAB; NOTINGHER, 2017).

Esse pré-processamento remove e minimiza as variações sistemáticas das amostras (BEEBE; PELL; SEASHOLTZ, 1998). Segundo Gemperline (2006) a normalização pode ser obtida pelas seguintes informações:

- Normalização pela área, onde divide cada variável por uma constante, deixando as variáveis em uma mesma escala, isto é, a constante é a soma dos valores absolutos das variáveis de uma determinada amostra;
- Normalização pelo mínimo e máximo, onde uma transformação linear é aplicada aos valores referentes a resposta Y no intervalo de mínimo e máximo (0 e 1);
- Normalização vetorial (NVU- do inglês, *Unit Vector Normalization*) divide os valores de cada número ou comprimento de onda pela norma do espectro, fazendo

com que ele seja um vetor unitário e assim conseguindo fazer a correção do espectro.

MSC

A correção do espalhamento multiplicativo (MSC – do inglês, *Multiplicative Signal Correction*) é um pré-processamento muito utilizado, o qual usa a regressão de cada espectro contra um espectro médio do conjunto de calibração (espectro de referencia). Portanto, a correção dos espectros é feita pela subtração dos valores originais o coeficiente linear da reta de regressão, sendo dividida pela inclinação da mesma regressão (NAES; MARTENS, 1989). Assim, a técnica calcula o espectro médio $\bar{x}$, para cada espectro individual x_i, e os coeficiente lienar (a_i) e angular (b_i) são estimados por regressão para corrigir cada espectros originais, conforme as equações abaixo (BORIN et al., 2007; MORGANO, 2005).

$$x_i = a_i + \bar{x} b_i \quad \text{Equação (1)}$$

$$x_{i\ corrigido} = \frac{x_i - a_i}{b_i} \quad \text{Equação (2)}$$

SNV

A variação normal padrão (SNV- do inglês, *Standard Normal Variate*) é um procedimento matemático aplicado para corrigir interferências de espalhamento e efeitos de espalhamento multiplicativo entre diferentes amostras (LUYPAERT, et al., 2004). A SNV é similar a MSC, porém são distintas. Por mais que ambas as técnicas tenham a finalidade de corrigir a linha de base, a SNV remove o espalhamento a partir da normalização de cada espectro pelo desvio padrão, isto é, padroniza cada espectro utilizando apenas os dados desse espectro (NAES; MARTENS, 1989). A SNV corrige o espectro individualmente, podendo ser aplicado quando há uma maior variabilidade entre um conjunto de espectros (COSTA FILHO, 2003). A correção realizada pela SNV é feita subtraindo o espectro inteiro $x_{i,j}$, pela média do mesmo espectro $\bar{x}_i$, divido pelo seu desvio padrão. Portanto, o SNV não precisa de um espectro "ideal" para realizar a correção dos espectros, Equação 3 e 4 (LUYPAET et al., 2004; COSTA FILHO, 2003).

$$\text{Média do espectro: } x = \sum_{j=1}^{n} \frac{X_{i,j}}{p} \quad \text{Equação (3)}$$

Espectro corrigido: $x_{i(SNV)} = \frac{(X_{i,j} - \bar{x}_i)}{\sqrt{\frac{\sum_{j=1}^{p} (X_{i,j} - \bar{x}_i)^2}{p-1}}}$ Equação (4)

Onde, x representa uma matriz (nxp) com espectros e comprimento de onda, $\bar{x}_i$ é a média do vetor contendo o espectro i da matriz x.

Correção de linha de base

Métodos de ajuste polinomial são simples e eficientes, podendo também serem aplicados para correção de linha de base e para remoção de interferências de fundo. O método de correção de linha de base por Mínimos Quadrados Ponderados (WLS, do inglês *Weighted least squares*) utiliza a variável que é referente a linha de base procedente do fundo e ajusta uma linha de base para cada espectro, identificando o sinal analítico e determinando quais vaiáveis estão acima da linha de base (CHEN et al., 2006)

Como as medidas de espalhamento Raman são afetadas por efeitos multiplicativos, aditivos e pela curvatura da linha de base, ocasionada pela fluorescência, o algoritmo WLS é eficiente para corrigir essas variações. Esse algoritmo tem uma linha de base de referência, fazendo o ajuste frequentemente aos pontos vizinhos do espectro que estão acima ou abaixo dessa linha de base de referência. E para isso polinômios de ordem 2 ou 3 são mais utilizados para organizar a linha de referência, devido o comportamento da variação de linha de base do espectro ocasionado pela fluorescência (CHEN et al., 2006).

Derivação

A derivada é um dos pré-processamento mais utilizados. Esse método consegue reduzir o efeito do espalhamento da luz, removendo os efeitos aditivos que aparecem aos espectros. A primeira e segunda derivada ressalta os picos e a relação entre eles. A primeira remove deslocamentos constantes da linha de base e a segunda elimina a variação linear da linha de base (FERRÃO, 2000; NAES *et al.*, 2002; TAYLOR; FRANCIS, 2006). Conforme a morfologia das partículas pode aparecer inclinação na linha de base dos espectros. Assim, a primeira e segunda derivada podem reduzir esses efeitos dos dados espectrais, fazendo com que as informações dos diferentes

comprimentos de onda sejam evidenciadas (FERRÃO, 2000; PIZARRO; ESTEBAN-DIEZ; GONZÁLEZ-SAIZ, 2004; MORGANO, 2005). Uma desvantagem vista nesse procedimento é que a presença dos ruídos nos espectros originais pode ser mais proeminente, e isso diminui a relação sinal/ruído (CHEN et al., 2006). Por outro lado, esse tipo de processamento beneficia problemas de sobreposição espectral, isolando bandas que são distintas e destacando estrutura fina dos espectros originais (LUYPAERT et al., 2004). Vários algoritmos são usados para o cálculo das derivadas, sendo a derivada de Savtzky-Golay a mais empregada para retirar ruídos que não são importantes (SOUSA; 2005).

Suavização

A suavização é a mais comumente utilizada para dados espectroscópicos, pois reduz o ruído de fundo mantendo as características de interesse. Para o uso desse processamento escolhe o número de pontos a serem suavizados (STEVENS; RAMIREZ-LOPEZ, 2017). Essa técnica de pré-processamento é expressa pela seguinte equação:

$$S_i = \frac{1}{\mathrm{M}} \sum_{j=0}^{M-1} x[i+j] \qquad \text{Equação (5)}$$

Onde, x é o espectro original (i= 1,2,3,...), S_i é o espectro depois da aplicação da suavização e M é o número de pontos utilizado na média móvel.

A suavização pelo filtro Savtzky Golay ajusta um polinômio em uma região simétrica em torno do ponto de interesse. O valor desse ponto é substituído pelo valor previsto dado pela função ajustada (RINNAN; VANDENBERG; ENGELSEN, 2009). Os filtros Savitzky Golay atuam sobre o sinal analítico ao inves da frequencia que compõem o sinal. Este procedimento envolve a derivada matemática de uma função, isto é a função é o espectro de uma detreminada amostra em vários comprimentos de onda. Este tipo de filtro é aplicado a sinais analíticos que apresentam picos estreitos, podendo aparecer distorções quando o número de pontos de intervalo é aumentado. Assim , o filtro SG se torna seletivo, ou seja, incapaz de diferenciar ruído e sinal, podendo remover informações importantes contida no sinal no processo de suavização de ruído (OTTO, 2007).

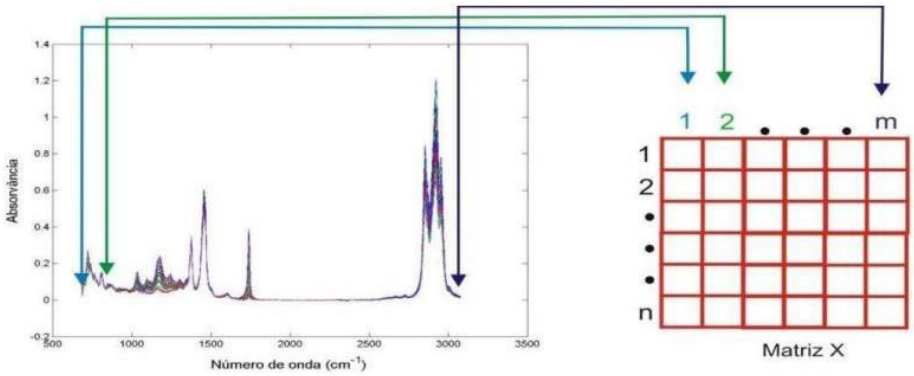
3.4.1 Construção de modelos multivariados

No ramo da análise multivariada a calibração se destaca. A calibração resulta na obtenção e validação de um modelo matemático, relacionando à informação obtida por um equipamento com a propriedade da amostra (GEMPERLINE, 2006). Os métodos de calibração multivariada podem ser associados às análises espectrofotométricas de multicomponentes, permitindo determinar parâmetros físico-químicos ou concentração de vários analitos ao mesmo tempo em uma amostra (ARAUJO *et al.*, 2001). A calibração multivariada permite determinar propriedades de vários analitos simultaneamente em uma amostra, estabelecendo relação entre a resposta do instrumento e a propriedade de interesse. (BEEBE; PELL; SEASHOLTZ, 1998).

Segundo Johnson e Wichern (1992), técnicas de análise multivariada podem ser aplicadas tanto para predição quanto para análise de relação entre variáveis. Dentre os métodos mais empregados para construção de modelos estão a Regressão Linear Múltipla (MLR), Regressão por Componentes Principais (PCR) e Regressão por Mínimos Quadrados Parciais (PLS) (MARTENS; NAES, 1996).

Para a construção de um modelo de calibração multivariada o conjunto de dados é dividido em conjunto de calibração, validação e predição. A matriz X corresponde as variáveis independentes relacionados ao sinal instrumental, como exemplo os valores de absorbância dos espectros. O vetor **Y** indica variáveis dependentes, referente as propriedades de interesse, como exemplo, as concentrações das amostras definidas pelo método de referência (BRAGA; POPPI, 2004). Assim, a matriz **X** é formada por *n* linhas (espectros) e *m* colunas (comprimento de ondas diferentes). A matriz **Y** também pode ser construída, por *n* linhas (valores de concentrações de diferentes amostras) e *q* colunas (número de diferentes analitos presentes na amostra) (NAGATA; BUENO; PERALTA-ZAMORA, 2001).

Figura 2 - Organização dos dados para calibração multivariada



Fonte: Cruz (2018).

3.4.1.1 Regressão por Mínimos Quadrados Parciais (PLS)

O método da regressão por mínimos quadrados parciais (PLS, do inglês "*Partial Least Squares*") trabalha os dados multivariados de forma a descobrir maior covariância entre a matriz de dados espectrais e o vetor **Y** (VANDEGINSTE et al., 1988). O PLS é o método de regressão que consiste na retirada de informações de absorbância do conjunto de dados da matriz **X**, correlacionando com informações retiradas de conjunto de dados de referência (vetor **Y**) a partir de operações matemáticas. Assim, a obtenção das variáveis latentes (VL) é feita pela combinação da matriz **X** com o vetor **Y**. A variável latente tem como objetivo alcançar a máxima covariância entre as concentrações das espécies e os espectros. Para construção do modelo de calibração um número de variável latente é usado de maneira que favoreça um menor erro de previsão, isto é, a diferença entre os valores previstos pelo modelo e valores de referência sejam o menor possível (WOLD; SJÖSTRÖM; ERIKSSON, 2001).

A técnica de calibração dos modelos PLS é a regressão mais utilizada a partir de dados de primeira ordem, podendo realizar a previsão das amostras não necessitando conhecer todos os componentes presentes nas amostras (BEEBE; PELL; SEASHOLTZ, 1998). Após a etapa de calibração, modelos de validação são construídos para um conjunto de amostras, onde as variáveis independentes obtidas são utilizadas em conjunto com o coeficiente de regressão para o cálculo de valores previstos para a variável

dependente. No processo de validação (interna e externa) as amostras das quais as variáveis dependentes são conhecidas para estabelecer uma comparação entre os valores previstos pelo modelo e valores conhecidos previamente a partir de uma metodologia padrão (HAALAND; THOMAS, 1988). Portanto, a regressão PLS decompõe a matriz **X** e **Y** em um conjunto de fatores ortogonais comuns e em um conjunto de pesos específicos. Isto é demostrado nas equações seguintes (BRUNS; FAIGLES, 1985).

$$X = T.P^T \qquad \textbf{Equação (6)}$$

Em que T é a matriz de score e P é a matriz de pesos.

$$Y = U.C^T \qquad \textbf{Equação (7)}$$

Em que U é a matriz de score e C é a matriz de pesos.

Assim, estabelece uma relação entre a matriz de scores de X e Y, através de uma regressão linear (Equação 8). Desta maneira, a matriz **U** (scores **Y**) podem ser previstos a partir da matriz T dos escores de X, e assim a variável de interesse pode ser prevista por meio dos dados espectrais contidos em X. Portanto, os métodos de calibração multivariada permitem a análise com a presença de interferentes, desde que estejam presentes na amostra utilizada para a construção do modelo de calibração. Assim, a calibração multivariada relaciona mais de uma resposta instrumental com a propriedade de interesse (VANDEGINSTE *et al.,* 1988).

$$\boldsymbol{U = bT + e} \qquad \text{Equação (8)}$$

3.4.1.2 Análise Discriminante por Mínimos Quadrados Parciais – PLS-DA

A análise discriminante por mínimos quadrados parciais é uma técnica de classificação linear que combina as características de regressão de mínimos quadrados parciais com o poder de discriminação. Este método é baseado na regressão PLS e procura umas variáveis latentes com uma covariância máxima com as variáveis Y. Quando um modelo PLS-DA é desenvolvido, é importante selecionar o número ideal de variáveis latentes (LV) e, geralmente, é feita com base na validação cruzada. A validação cruzada é realizada a partir da divisão de amostras em grupos de validação cruzada, isto é, cada grupo é retirado do conjunto de treinamento uma por vez, sendo comum quando o número de amostras é pequeno. Assim, o modelo é calibrado com as amostras de treinamento restantes e logo após são usadas para prever as amostras do grupo de validação cruzada (BALLABIO; CONSONNI,2013).

No PLS-DA é estabelecida uma relação linear entre a variável dependente (Y) e as variáveis independentes (**X**). Tanto o PLS-DA quanto o PLS são baseados na análise de componentes principais (PCA), ou seja, a matriz **X** é decomposta em duas matrizes, scores e loadings, buscando a máxima covariância entre as matrizes de **X** com **Y**. No modelo PLS-DA a matriz Y contém informações sobre as classes das amostras e no PLS a matriz **Y** contém os valores de informações de uma propriedade de interesse (FERREIRA, 2015).

O desenvolvimento do modelo consiste em duas etapas: o treinamento/calibração e o teste/validação. A primeira etapa investiga as propriedades dos dados, disposto a encontrar um modelo para o seu comportamento. Já na segunda etapa algumas amostras que não participam da calibração são usadas para avaliar a qualidade do modelo construído (GELADI; KOWALSKI, 1986).

O PLS-DA é aplicado neste trabalho com o intuito de correlacionar os dados espectroscópicos Raman (matriz **X**) com os dados de composição (matriz **Y**) para a construção de um modelo de classificação, permitindo a previsão de componentes desejados a partir de um espectro medido.

3.5 Figuras de Mérito

O desenvolvimento de um método analítico requer a avaliação e verificação da confiabilidade, no qual, é regulamentada e orientada por órgãos de fiscalização, como o Instituto Nacional de Metrologia, Normalização e Qualidade Industrial (INMETRO) e a Agência Nacional de Vigilância Sanitária (ANVISA), demonstrando a eficiência do método sendo adequado para determinado propósito, em que esse processo é denominado validação. O processo de validação é importante para verificar se um método satisfaz o seu objetivo conforme o nível de confiança estatística. Os cálculos dos parâmetros de validação são feitos baseados na calibração entre um sinal analítico, referente ao analito, e o valor de referência. Todavia, o sinal analítico é pertencente tanto à matriz quanto ao analito, isto é, sendo referente a análise multivariada (WILLIAMS; SWINKELS; MAEDER, 1992; SILVA, 2012; SILVA; FERREIRA; BRAGA, 2021).

Lorber definiu o Sinal Analítico Líquido como sendo a parte do sinal analítico que provém do analito, isto é, ortogonal às subvenções de outros possíveis constituintes presentes na amostra. Inicialmente, o método de NAS foi proposto para modelos de mínimos quadrados clássicos (CLS – *classical least squares*), e posteriormente para uma

aplicação mais ampla em calibração multivariada modificou-se o método e, assim, sendo possível utilizá-lo em métodos de PCR e PLS (LORBER, 1986).

Um método de calibração multivariada robusto precisa ser calibrado com amostras representativas. E para isso, são feitas remoção de amostras anômalas (outliers) e seleção de amostras de calibração. Em que são consideradas amostras anômalas quando apresentam um alto valor de influência no modelo (laverage) e, portanto, não sendo bem modeladas e previstas (WILLIAMS; SWINKELS; MAEDER, 1992; SILVA, 2012).

O desempenho dos modelos de classificação pode ser avaliado por diferentes parâmetros que são baseados em possíveis respostas binárias, como: verdadeiro positivo (VP – o modelo classifica como positiva quando a amostra é realmente positiva), falso positivo (FP – o modelo classifica como positivo uma amostra negativa), verdadeiro negativo (VN - o modelo classifica como negativa quando a amostra é realmente negativa), e o falso negativo (FN – o modelo classifica a amostra como negativa, quando esta é positiva) (VALENTINI; SOMMER; MATIOLI, 2007; MORAIS, 2014).

A avaliação dos modelos de calibração é feita a partir de figuras de mérito multivariadas, como forma de julgar o desempenho dos métodos quimiométricos. Entre algumas delas estão: Sensibilidade Analítica, Raiz Quadrada do Erro Médio Quadrático da Previsão (RMSEP), coeficiente de correlação (R2), Limite de Detecção (LOD), Raiz Quadrada do Erro Médio Quadrático da Validação Cruzada (RMSECV) e Seletividade/Especificidade Analítica (OLIVERI et al., 2006; OLIVERI; FABER, 2009).

Seletividade/Especificidade analítica

Na validação a seletividade é o primeiro método a ser analisado, e indica a medida do grau de sobreposição entre o sinal da espécie de interesse e os interferentes presentes na amostra (VESSMAN et al., 2001). Assim, se um método não for seletivo, parâmetros como precisão, exatidão e linearidade podem se comprometerem, portanto, é importante que a seletividade seja executada continuamente durante o processo e validação do método (VALENTINI; SOMMER; MATIOLI, 2007). A seletividade/especificidade corresponde a fração de amostras que não pertencem à classe e são classificadas corretamente como não pertencentes à classe (OLIVERI; DOWNEY, 2012):

$$Especificidade = \frac{VN}{VN+FP} \quad \text{Equação (9)}$$

Sensibilidade Analítica

A sensibilidade é a fração das amostras classificadas corretamente na classe modelada, sendo apta pelo modelo, medida pelo nível de confiança do espaço das classes (OLIVERI; DOWNEY,2012):

$$Sensibilidade = \frac{VP}{VP+FN} \quad \text{Equação (10)}$$

Raiz Quadrada do Erro Médio Quadrático da Previsão (RMSEP) e da Validação Cruzada (RMSECV)

A raiz quadrada do erro médio quadrático da previsão (RMSEP) e de calibração (RMSEC) é expresso através da exatidão que demonstra o grau de concordância entre o valor estimado pelo modelo multivariado e o valor tido como verdadeiro obtido pelo método de referência. (BRERETON, 2000).

A avaliação da exatidão de um modelo pode ser feita conforme as equações (9) e (10) (FERREIRA, 2015):

$$RMSEP = \frac{\sqrt{\sum_{i=1}^{I} (yi - \hat{y}i)^2}}{I} \quad \text{Equação (11)}$$

$$RMSECV = \frac{\sqrt{\sum_{i=1}^{I} (yi - \hat{y}i)^2}}{I - A} \quad \text{Equação (12)}$$

Onde,

Yi = valor de referência

Ŷi = valor previsto da amostra

I = número de amostras na validação cruzada

A = variáveis latentes

Curva ROC

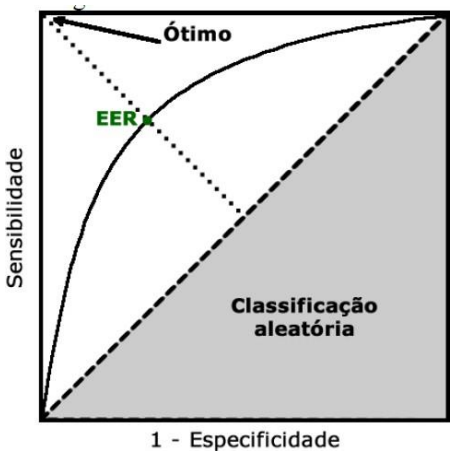
Para modelos de classificação é possível usar a área abaixo da curva (AUC) de características de operação do receptor (ROC - *Receiver Operating Characteristic*) permite uma melhor visualização de informações (BRERETON, 2018). A área sob a

curva representa o desempenho de um algoritmo no espaço ROC, e quanto mais próximo do valor unitário, melhor será o desempenho do classificador (GONZAGA, 2011).

Por sua vez, a curva ROC é uma ferramenta que avalia a conformidade e o desempenho de um modelo de classificação (GONZÁLEZG et al., 2014). Nesta curva os valores das figuras de mérito sensibilidade e especificidade são representadas a partir de um gráfico cartesiano que facilita a seleção de classificadores por seu desempenho (FAWCETT, 2006).

Para a análise da curva ROC é traçado uma curva sob os eixos cartesianos, indicando valores da fração de falsos positivos (especificidade) no eixo das abcissas e valores da fração de verdadeiros positivos (sensibilidade) no eixo das ordenadas. O ponto ideal é indicado pela seta mostrado na Figura 3 e quanto mais próximo à curva estiver deste ponto, maior será a capacidade de discriminação do modelo. A região triangular sombreada indica uma área em torno de 0,5 mostrando uma predição aleatória. A reta diagonal pontilhada mostra os valores de especificidade e sensibilidade são iguais, diante disto, o ponto da curva que cruza essa semi diagonal é a taxa de erro igual (ERR) (SOVIERZOSKI, 2009). Portanto, as figuras de mérito são indicadores quantitativos do escopo e do bom desempenho das técnicas.

Figura 3 – Curva ROC hipotética



Fonte: Sovierzoski (2009).

4 MATERIAL E MÉTODOS

As medidas espectrais foram realizadas no Laboratório Avançado de Tecnologia Química (LATECQ) da Embrapa Algodão, localizado na cidade de Campina Grande, estado da Paraíba. O tratamento dos dados obtidos foi realizado no Laboratório de Síntese Inorgânica e Quimiometria (LabSIQ/CCT/UEPB).

4.1 Amostras

As amostras das sementes de algodão convencionais e transgênicos foram da cultivares de algodão - BRS Aroeira (convencional) e a BRS 268 RF (transgênica). As sementes coletadas foram limpas, deixando-as sem resíduos de fibras, acondicionada em sala climatizada por 24 h a temperatura de 20^0C e umidade relativa de 65%.

Cada semente foi seccionada no sentido longitudinal em duas faces opostas para obtenção dos espectros. A construção dos modelos de classificação foi realizada com 20 amostras individuais de sementes de cada cultivar. Para cada semente foram obtidos seis espectros, sendo três repetições para cada lado da semente. Totalizando um conjunto de dados com 120 espectros. Devido à baixa qualidade e a relação sinal/ruído 20 espectros precisaram ser excluídos do conjunto de dados.

4.2. Instrumentação analítica

As medidas espectrais foram adquiridas por meio do espectrômetro Ramanportátil modelo Mira Cal DS da marca Metrohm (**Figura 4**), com faixa espectral de 400 a 2300 cm^{-1}, resolução espectral de 8-10 cm^{-1}, que possui um único laser de comprimentode onda de 785 nm ± 0.5nm e potência de aproximadamente 100 mW e interface USB 2.0 com fonte de alimentação de transmissão de dados com cabo USB.

Figura 4 - Espectrômetro Raman Portátil

Fonte: Metrohm.com (2020)

4.3 Tratamento dos dados

Os dados espectrais foram tratados utilizando o pacote PLS Toolbox, versão 6.0.1 da marca Eigenvector, escrito em MATLAB®.

5 RESULTADOS E DISCUSSÃO

5.1 Análise dos espectros RAMAN

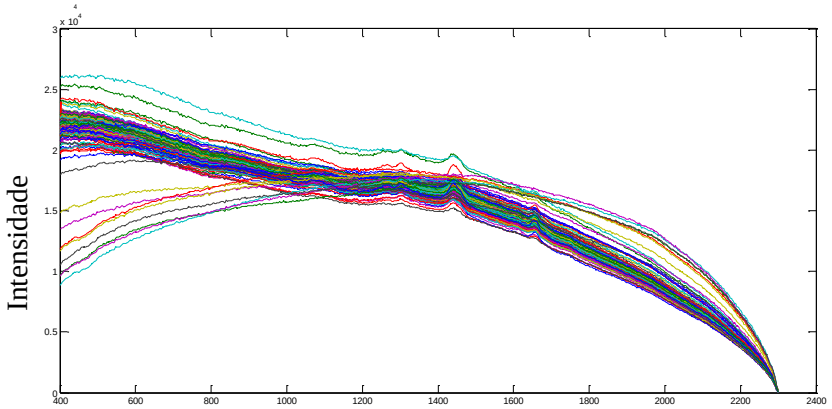
As medidas nas sementes foram processadas em duas condições para leitura efetiva. Inicialmente, nas sementes intactas e, depois, com as sementes cortadas que possíveis interferências do tegumento da semente fossem superadas, ou tivessem mínima interferência devido a coloração, de cor preta e opaca, ou ainda resíduos de línter (camada de fibras curtas aderidas à superfície das sementes) característico dessa matriz. Deste modo, os melhores resultados foram obtidos com as sementes seccionadas ao meio (**Figura 5**).

Figura 5 - Sementes de algodão convencionais e transgênicas.

Fonte: Própria, 2020.

Os espectros Raman obtidos na região de 400 a 2300 cm^{-1} de todas as sementes analisadas são mostrados na **Figura 6**. Pode-se notar uma curvatura nos extremos do espectro, influência do sinal de fluorescência, além de um deslocamento de linha de base e regiões com baixa relação sinal ruído (BURGER e GOWENE, 2011).

Figura 6 - Espectros brutos das sementes convencionais e transgênicas

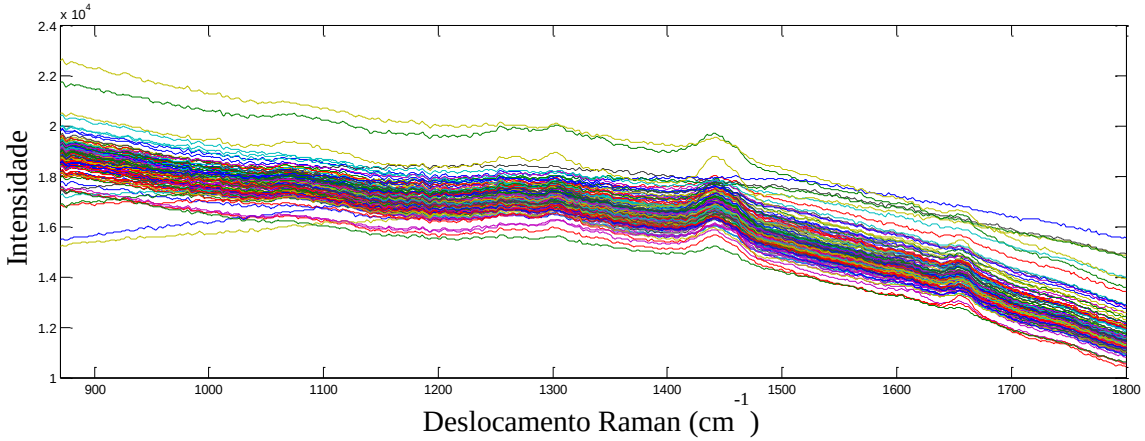


Deslocamento Raman (cm)

Fonte: Própria,2020.

Para um melhor desempenho das técnicas de pré-processamento as regiões externas que compreende a faixa espectral de 400 a 869 cm^{-1} e de 1801 cm^{-1} a 2300 cm^{-1} foram excluídas da base de dados. Nessas regiões excluídas observa-se umacurvatura que dificulta a aplicação das técnicas de pré-processamento espectral. Assim, aregião espectral trabalhada foi de 870 cm^{-1} até 1800 cm^{-1} (Figura 7), correspondente a uma matriz com dimensão de 120 x 931.

Figura 7 - Espectros brutos sem as regiões externas das sementes convencionais e transgênicas.

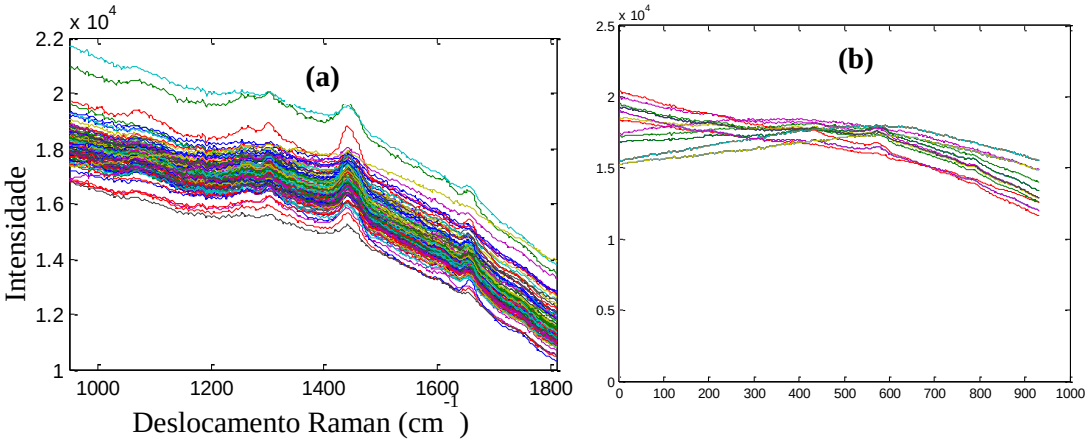


Fonte: Própria, 2021.

5.2 Pré-processamento dos dados Raman

Após essa análise prévia e antes da aplicação de técnicas de pré-processamento para a construção do modelo de calibração multivariada, foi realizado um estudo criterioso sobre o comportamento dos espectros e verificou-se que alguns desses espectros não apresentavam perfil semelhante ao conjunto dos dados espectrais (Figura 8a). Deste modo, se fez necessário remover do conjunto de dados 20 espectros que não apresentava um bom perfil espectral (Figura 8b).

Figura 8 – Espectros brutos com as 120 amostras (a) e espectros removidos do banco de dados (b).



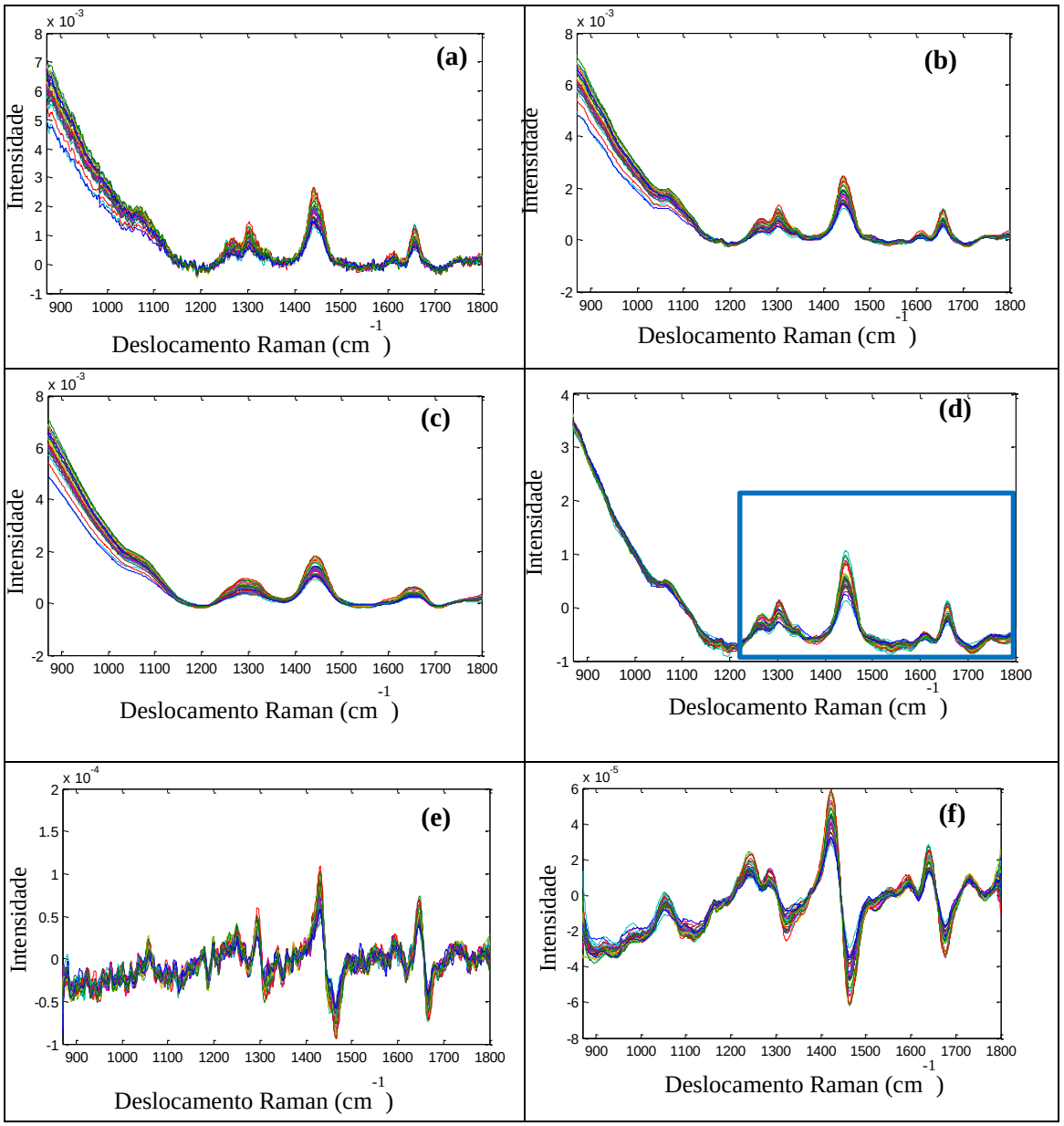
Fonte: Própria, 2021.

Posteriormente, ao fazer um estudo criterioso sobre o comportamento dos espectros, diversas técnicas de pré-processamento foram aplicadas a fim de verificar qual destas forneceria o melhor desempenho na construção do modelo PLS-DA.

Os efeitos físicos mais comuns nas medidas de espalhamento e reflectância Raman em materiais sólidos é o espalhamento da radiação, provocando deslocamento na linha de base dos espectros. Essa variação decorre do tamanho e forma das partículas, fluorescência, comprimento de onda incidente, índice de refração da amostra, entre outros elementos (Ferreira, 2015). Deste modo, verificou-se a necessidade do uso de técnicas de correção a fim de minimizar o efeito da fluorescência presente nos espectros Raman. Existem inúmeros algoritmos matemáticos que minimizam essa variação da linha de base, neste trabalho a normalização pelo vetor unitário (NVU), juntamente com a correção de linha de base com o algoritmo WLS (do inglês *Weight Least Squares*) foram aplicados como pre-processamento padrão e em conjunto com diversos outros pré-processamentos, afim de verificar qual obteria o melhor desempenho.

Os pré-processamentos adicionados ao preprocessamento padrão (NVU + WLS) foram (Figura 9a): Suavização Savitzky-Golay de 1ª ordem e janela de 15 pontos (figura 9b); Suavização Savitzky-Golay de 1ª ordem e janela de 51 pontos (figura 9c); Suavização Savitzky-Golay de 1ª ordem e janela de 15 pontos e Variação Normal Padrão (do inglês *Standard Normal Variate* – SNV) (figura 9d); 1ª derivada - polinômio de 1ª ordem e janela de 15 pontos (figura 9e); 2ª derivada - polinômio de 2ª ordem e janela de 51 pontos (figura 9f).

Figura 9 – Dados pré-processados com **a)** NVU+WLS; **b)** NVU+WLS +suavização com polinômio Savitzky-Golay de 1ª ordem e janela de 15 pontos; **c)** NVU+WLS +suavização com polinômio Savitzky-Golay de 1ª ordem e janela de 51; **d)** NVU+WLS +suavização com polinômio Savitzky-Golay de 1ª ordem e janela de 15 pontos + SNV ; **e)** NVU+WLS +1ª derivada com polinômio Savitzky-Golay de 1ª ordem e janela de 15 pontos ; e **f)** NVU+WLS +1ª derivada com polinômio Savitzky-Golay de 2ª ordem e janela de 51 pontos.



Fonte: Própria, 2020.

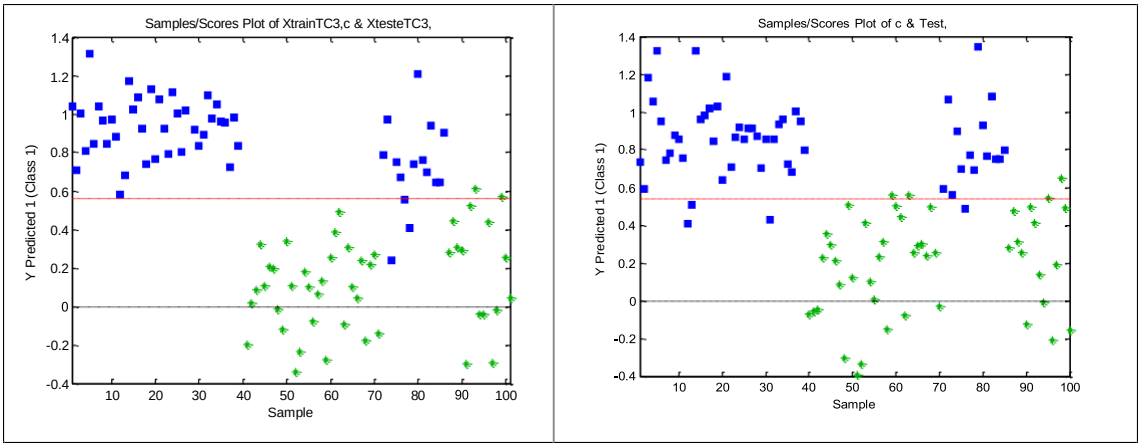
5.3 Modelo PLS-DA

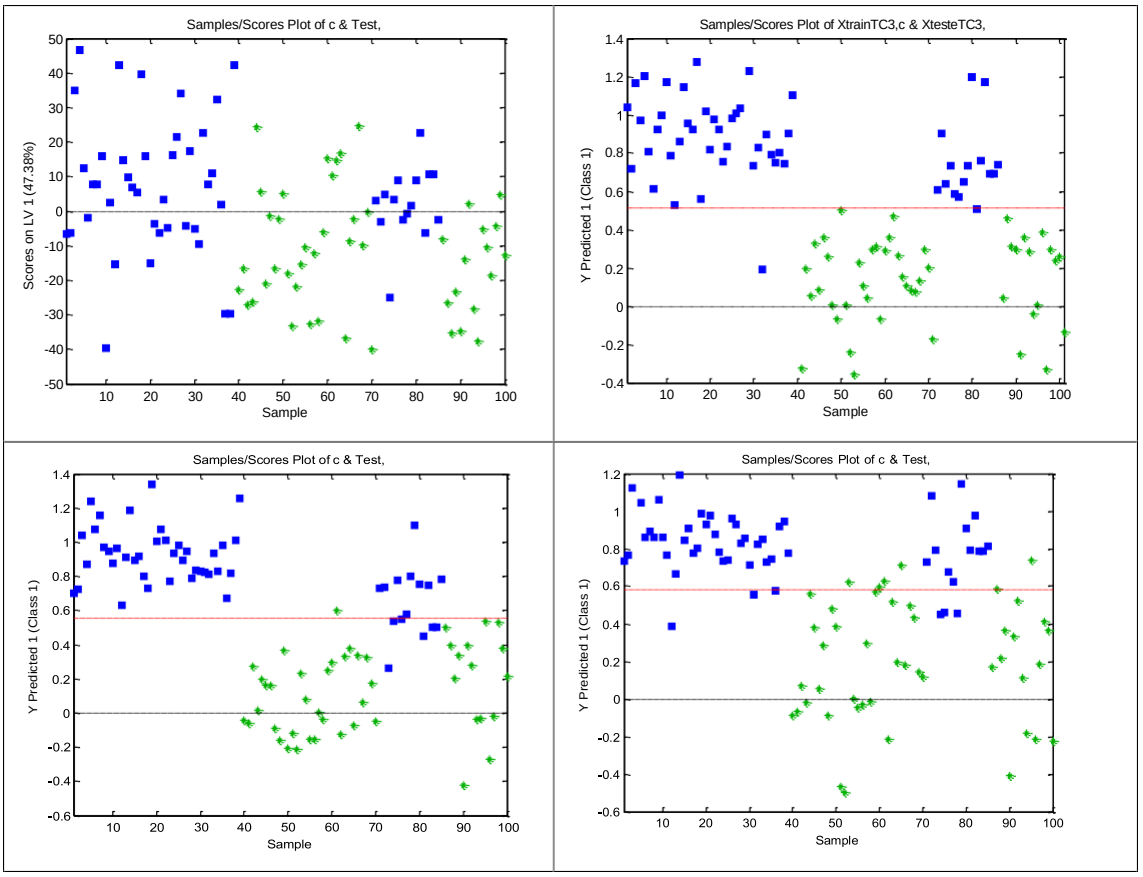
O modelo PLS-DA foi elaborado para distinguir entre as amostras de sementes convencionais e geneticamente modificadas. Para a construção dos modelos de reconhecimento de padrões supervisionados utilizando o PLS-DA selecionou-se para o conjunto de treinamento 70 espectros (~12 sementes) sendo 39 das sementes transgênicas e 31 das sementes convencionais. Já, para o conjunto de teste foram selecionados 30 espectros (5 sementes), sendo 15 de cada tipo.

Assim, os dados foram organizados de modo que os espectros de uma mesma semente estivesse em apenas em um dos subconjuntos. Ou seja, que não houvessem espectros da mesma semente na etapa de treinamento e de teste. Como vetor de atributo foi usado o número 0 (zero) para as amostras convencionais e 1 (um) para as amostras transgênicas (100 linhas por 1 coluna).

Para avaliar os modelos de classificação foram considerados dois parâmetros importantes, a sensibilidade e especificidade, que são calculadas de acordo com as equações 9 e 10. Também foi avaliado o gráfico de escores das amostras de teste (Figura 10) que mostra a distinção entre as classes convencionais e transgênica para cada pré-processamento em um limiar de 0,5. As amostras com coloração verde representam as sementes convencionais e as azuis as transgênicas.

Figura 10 – Gráfico dos escores para a predição das amostras tratadas com **a)** NVU+WLS, **b)** NVU+WLS +suavização com polinômio SG de 1ª ordem e janela de 15 pontos, **c)** NVU+WLS +suavização com polinômio SG de 1ª ordem e janela de 51**, d)** NVU+WLS +suavização com polinômio SG de 1ª ordem e janela de 15 pontos + SNV utilizando apenas a região destacada em anterior (Figura 9d); **e)** NVU+WLS +1ª derivada com polinômio SG de 1ª ordem e janela de 15 pontos; e **f)** NVU+WLS +2ª derivada com polinômio SG de 1ª ordem e janela de 51.





Fonte: Própria, 2020.

Pode-se observar que na Figura 10b o modelo conseguiu diferenciar as amostras que não pertencem a classe na etapa de treinamento e classificar as amostras pertencentes a uma classe na etapa de teste. O modelo da Figura 10c foi mais eficaz em classificar as amostras pertencentes a uma classe (sensibilidade) na etapa de treinamento, do que as diferenciar das amostras que não pertencem aquela classe. O mesmo comportamento foi observado na etapa de teste. Na Figura 10d o respectivo modelo conseguiu diferenciar as amostras que não pertencem a uma classe, tanto na etapa de treinamento quanto na etapa de teste Já, na Figura 10e e 10f o modelo conseguiu classificar as amostras pertencentes aquela classe na etapa de treinamento e diferenciar as amostras que não pertencem na etapa de teste. Isto é, a variância da performance da sua capacidade de generalização não é boa. Assim, os modelos não conseguiram realizar estimativas precisas. E isto pode ser visto nos valores das figuras de mérito de cada pré-processamento (Tabela 1). Diante do dos valores das figuras de mérito apresentadas na tabela 1 o melhor modelo foi construído com toda faixa espectral e utilizando os pré-processamentos **NVU+WLS +S15+SNV**.

Tabela 1 – Figuras de mérito de cada pré-processamento

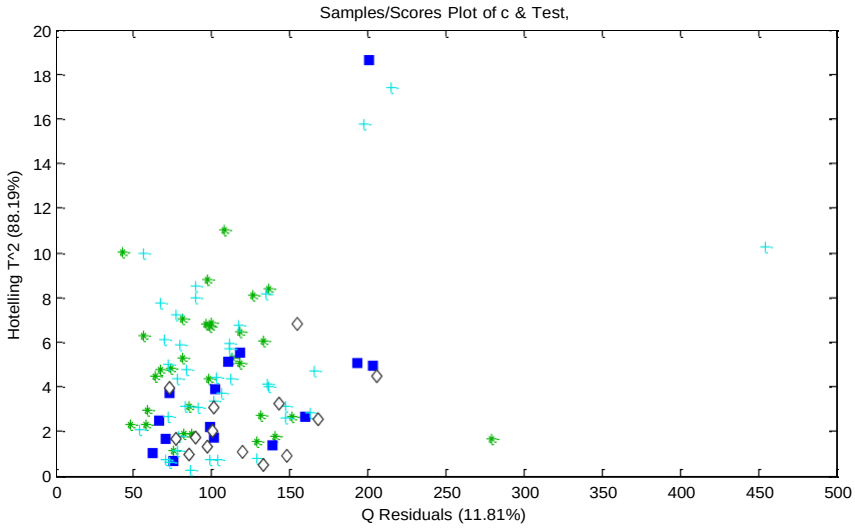
Figuras de mérito

Pré-processamento	VL	Treinamento		Teste		Treinamento	Teste	R^2
		Sensibilidade	Especificidade	Sensibilidade	Especificidade	Erro	Erro	
NVU+WLS	5	1,00	1,00	0,800	0,867	0,000	0,170	0,560
NVU+WLS +S15	6	0,923	0,935	0,933	0,867	0,071	0,100	0,575
NVU+WLS +S51	7	0,949	0,871	1,00	0,867	0,090	0,066	0,569
NVU+WLS +S15+SNV	**5**	**0,974**	**0,968**	**0,933**	**1,00**	**0,029**	**0,033**	**0,643**
NVU+WLS +1DERJ15	3	1,00	0,968	0,600	1,00	0,016	0,20	0,465
NVU+WLS +S15+SNV(faixa)	6	0,923	0,968	0,667	0,867	0,054	0,233	0,371
NVU+WLS +1DER2J51	4	0,923	0,871	0,800	0,867	0,103	0,166	0,525

Onde: NVU: Normalização pelo vetor unitário; WLS: Correção de linha de base; S15 : Suavização com polinômio Savitzky-Golay de 1ª ordem e janela de 15 pontos e S51: Suavização com polinômio Savitzky-Golay de 1ª ordem e janela de 51; SNV: , 1DERJ15: 1ª derivada com polinômio Savitzky-Golay de 1ª ordem e janela de 15 pontos; e 1DER2J51: 1ª derivada com polinômio Savitzky-Golay de 2ª ordem e janela de 51.

A avaliação sobre a presença de amostras anômalas foi realizada com base no gráfico de T^2 de Hotelling versus $Q_{Residuals}$, mostrado na figura 11.

Figura 11 – T^2 Hotteling versus Q Residuals

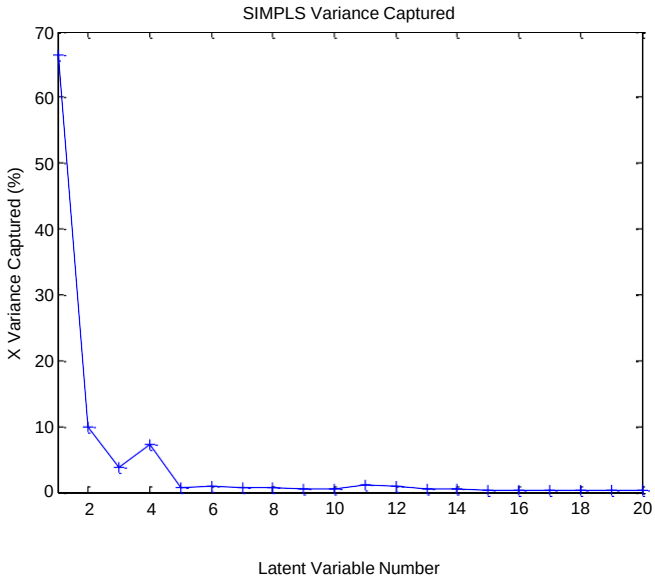


Fonte: Própria, 2020.

Pode-se observar que algumas amostras apresentaram alto valor de T^2 Hotelling ou de $Q_{Residuals}$. No entanto, essas não foram consideradas anômalas por não por não possuírem valores desses dois parâmetros elevados de forma concomitante.

O modelo foi construído com 5 variáveis latentes, que explicam 88,19% da variância dos dados (Figura 12).

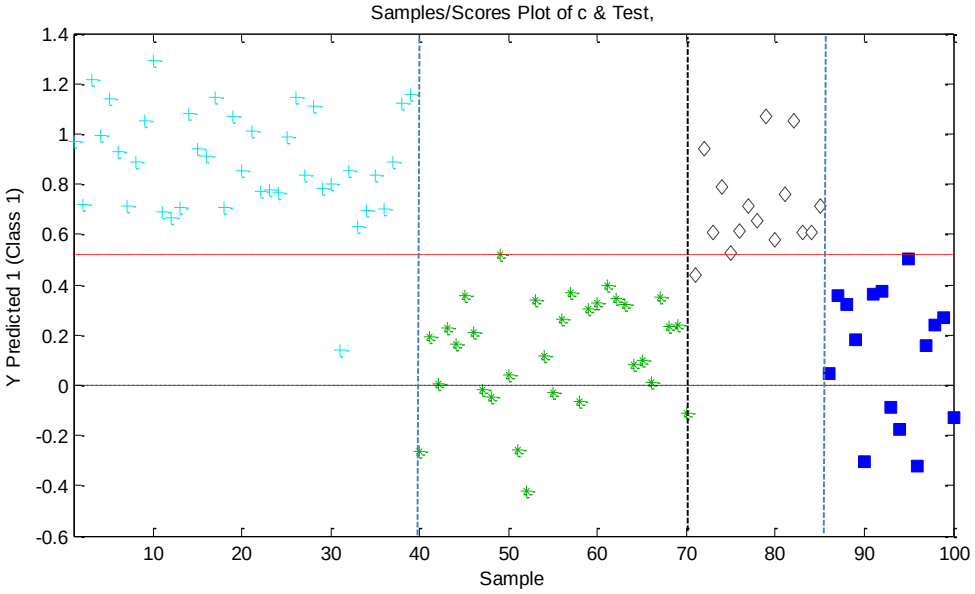
Figura 12 – Número de Variáveis Latentes.



Fonte: Própria, 2020.

A Figura 13 mostra a classificação de sementes de algodão convencionais e geneticamente modificadas pelo modelo PLS-DA (NVU+WLS +S15+SNV) que apresentou os melhores valores de figuras de mérito.

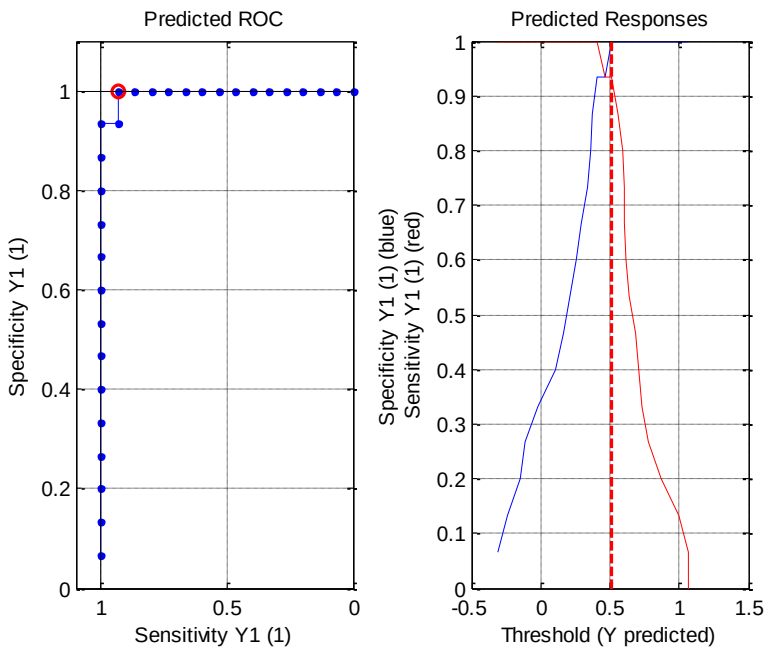
Figura 13 – Predição das classes



Fonte: Própria, 2020.

Nela as amostras geneticamente modificadas (GM) são representadas em azul claro e as convencionais em verde, na etapa de treinamento. Na etapa de teste, as amostras GM são representadas por losangos brancos e as convencionais por quadrados azuis. O modelo apresentou uma sensibilidade (amostras pertencentes à classe e classificadas corretamente nesta classe) e especificidade (amostras não pertencentes à classe modelada e corretamente classificados como não pertencentes) de 0,974 e 0,968 na etapa de treinamento e 0,933 e 1,0 na etapa de teste, respectivamente, como mostrado na curva ROC apresentada na Figura 14. Ela mostra o limite entre as classes determinado a partir das previsões do conjunto de treinamento para a discriminação correta de todas as amostras no conjunto teste. Os valores de sensibilidade e seletividade para um modelo ideal devem ser próximos ao valor unitário. A curva ROC indica que esse modelo é perfeitamente capaz de discriminar as classes de sementes convencionais e transgênicas.

Figura 14 – Curva ROC do modelo PLS-DA

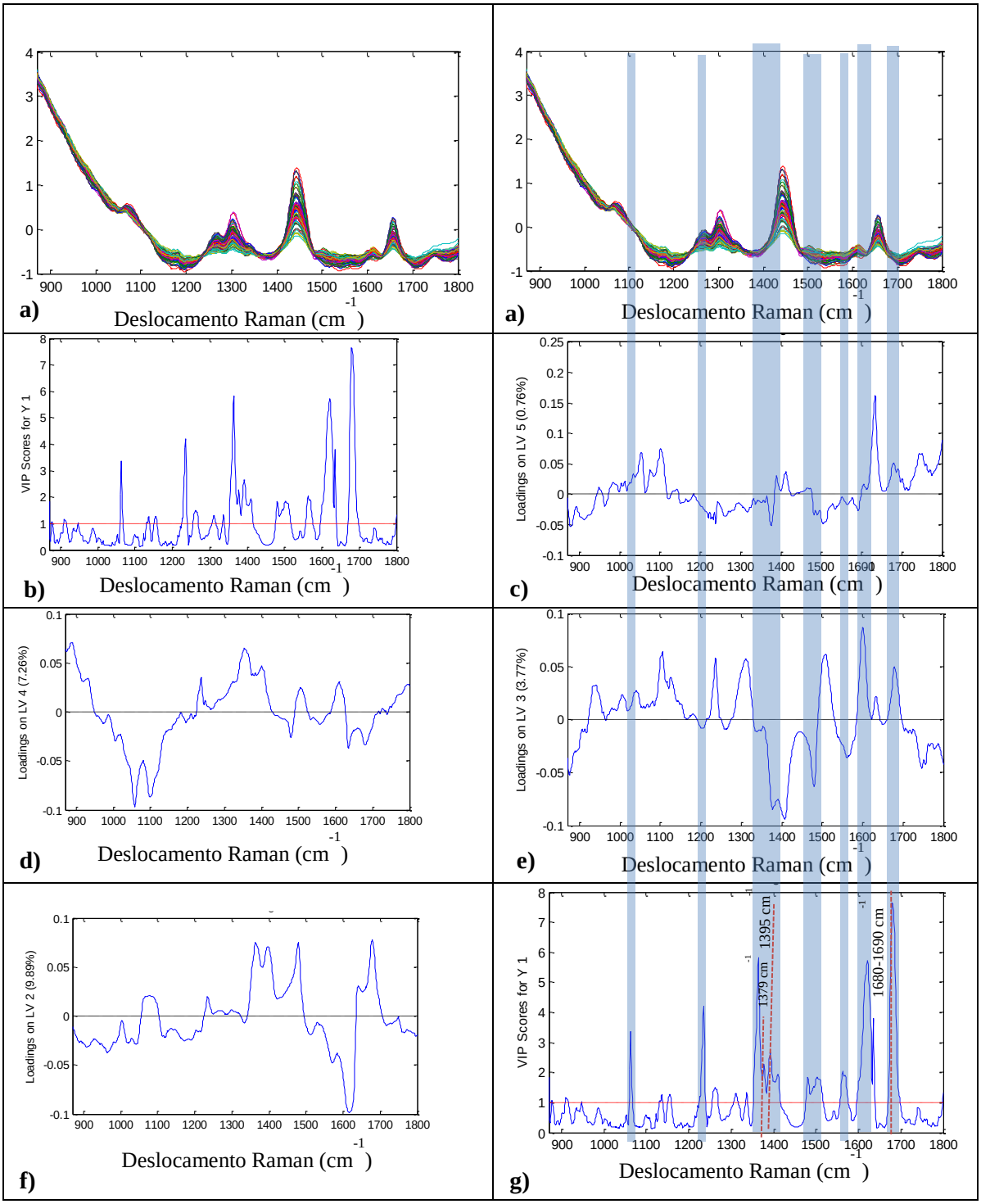


Fonte: Própria, 2020.

O bom desempenho do modelo é expresso também pelos valores de erro de classificação obtidos na etapa de treinamento e de teste, que foram de 2,9% e 3.3%, respectivamente. Além disso, é possível observar uma maior dispersão entre as amostras convencionais que as amostras de sementes transgênicas, tal como valores inferiores de erro de classificação por modelos construídos por espectroscopia NIR (ROCHA, 2020).

A Figura 15 mostra os espectros das sementes de algodão (a), o gráfico de loadings para as 5 primeiras variáveis latentes (b até f) e os escores VIP´s (g), nos quais pode-se observar as variáveis espectrais que têm contribuição para a distinção entre as sementes convencionais e transgênicas.

Figura 15 – a) Espectros pré-processados, b) até f) loadings nas variáveis latentes de 1 a 5, respectivamente e g) escores VIP´s. As faixas destacadas em azul são as regiões importantes de acordo com o gráfico de VIPs escores.



Fonte: Própria, 2020.

De acordo com a Figura 15, podemos observar que as 5 variáveis latentes trazem informação espectral para o modelo. De acordo com os escores VIPs, as regiões espectrais destacadas em azul, são variáveis relacionadas ao problema estudado e que são importantes para a descriminação entre as sementes geneticamente modificadas e as convencionais. O gráfico dos VIP´s escores mostram as variáveis espectrais que mais contribuem para a separação entre as classes estudadas. Nele podemos observar o limiar (tracejado em vermelho), onde as variáveis com valores superiores a 1 representam as regiões espectrais mais significativas.

Assim, com base na literatura (RYGULA, et. al, 2013; RON; JOHN, 2016; EDWARDS; FARWELL; WEBSTER, 1997) tentou-se fazer a atribuição de algumas bandas. A banda larga em torno de 1680-1690 cm^{-1} corresponde a amida I (RYGULA, et. al, 2013). A banda em torno de 1396 cm^{-1} corresponde ao estiramento do grupo C-C-H do carboidrato (BENZERDJEB; MOKHTARI; RAHAL, 2007) e o comprimento de onda de 1379 cm^{-1} é referente ao estiramento CH2 (EDWARDS; FARWELL; WEBSTER, 1997) que podem ser provenientes das moléculas lipídicas presente na composição do caroço de algodão.

6 CONCLUSÃO

A aplicação da tecnologia portátil que emprega espectroscopia Raman e o método de reconhecimento de padrões (PLS-DA) para a distinção entre cultivares convencionais e geneticamente modificadas de algodão. Foi verificado que o tipo de pré-processamento, bem como a região espectral utilizada, é determinante para a melhoria do desempenho dos modelos construídos. Neste sentido, o modelo PLS-DA (NVU+WLS +S15+SNV) forneceu os melhores valores de especificidade, sensibilidade e erro de classificação nas etapas de treinamento e teste.

A utilização do modelo PLS-DA proporcionou bons resultados, demostrando sua potencialidade na construção de modelos para identificação de sementes convencionais e transgênicas de algodão. Usando apenas 5VLs, o erro obtido para o conjunto de amostras foi de 3,3% na etapa de predição e de 2,9% na etapa de calibração. A etapa de treinamento apresentou uma especificidade de 0,968 e sensibilidade de 0,974, mostrando um alto poder de distinção. Já, na etapa de teste o modelo apresentou uma especificidade de 1,00 e uma sensibilidade de 0,933, mostrando uma boa classificação entre as classes modeladas.

Os resultados demonstraram a importância na escolha e no tipo de pré-processamento empregado no tratamento de dados. Apesar de ser um método destrutivo, os valores de erro de classificação são bastante inferiores aos obtidos por modelos construídos utilizando a espectroscopia NIR.

O método se apresenta como uma importante prova de teste para o uso do espectrômetro Raman portátil na identificação de sementes GM de algodão. E pode ser vista como uma alternativa vantajosa para identificação de sementes de algodão convencional e transgênicas.

REFERÊNCIAS

ABBAS, O.; DARDENNE, P.; BAETEN, V. **Near-Infrared, Mid-Infrared, and Raman Spectroscopy.** Chemical Analysis of Food: Techniques and Applications, 2012.

ABBAS, Q; AHMAD, S. **Effect of different sowing times and cultivars on cotton fiber quality under stable cotton-wheat cropping system in southern Punjab.** Pakistan Journal of Life and Social Sciences. v. 16, p. 77–84, 2018.

AHMAD, M. M.; ALI, A.; SDDIQUI, S.; KAMALUDDIN; ABDIN, M. Z. **Methods in Transgenic Technology**. Plant Biotechnology: Principles and Applications, 2017.

AHMAD, S.; RAZA, I. **Optimization of management practices to improve cotton fiber quality under irrigated arid environment.** Journal Food Agriculture Environment, v. 2, n, 2.p. 609–613, 2014.

ALMEIDA, M. R.; CORREA, D. N.; ZACCA, J. J.; LOGRADO. L. P; POPPI, R. J. **Detection of explosives on the surface of banknotes by Raman hyperspectral imaging and independent component analysis**. Analytica Chimica Acta ,v. 860, p. 15-22, 2015.

ARAÚJO, M. C. U. *et al.* **The successive projections algorithm for variable selection in spectroscopic multicomponent analysis**. Chemometrics and Intelligent Laboratory Systems, v. 57, n. 2, p. 65 – 73, 2001.

BAGATIN, E. **Envelhecimento cutâneo e o papel dos cosmecêuticos.** São Paulo. Boletim Dermatológico UNIFESP, v. 5, n.17, p. 1- 4, 2008.

BALLABIO, D.; CONSONNI, V. allabio, D. **Classification tools in chemistry. Part 1: linear models. PLS-DA** Analytical Methods , v.*5*, p. 3790, 2013.

BARANSKI R. ; BARANSKA, M. **Discrimination between Nongenetically Modified (Non-GM) and GM Plant Tissue Expressing Cysteine-Rich Polypeptide Using FT-Raman Spectroscopy**. J. Agric. Food Chem., Vol. 56, No. 12, 2008

BARROSO, P. A. V.; FREIRE, E. C.; AMARAL, J. A. B.; SILVA, M. T. **Zonas de exclusão de algodoeiros transgênicos para preservação de espécie *Gossypym* nativas ou naturais**. Comunicado técnico n. 242, agosto/ 2005

BEEBE, K. R.; PELL, R. J.; SEASHOLTZ, M. B. **Chemometrics: a Pratical Guide.** 1ª Ed. Ed. A John Wiley & Sons. New York (1998).

BENZERDJEB, A. M.; MOKHTARI, I. N. T.; RAHAL, M. S. **Normal coordinates analyses of disaccharides constituted by D-glucose, D-galactose and D-fructose units**. Spectrochimica Acta Part A, v.68, p. 284–299, 2007.

BOCKLITZ, T.; WALTER, A.; HARTMANN, K.; RÖSCH, P.; POPP, J. **How to pre-process Raman spectra for reliable and stable models?** Analytica Chimica Acta, v. 704, p. 47–56, 2011.

BOBROWSKI, V.L; FIUZA, L. M.; PASQUALI, G.; BODANESE-ZANETTINI, M.H. **Genes de Bacillus thuringiensis: uma estratégia para conferir resistência a insetos em plantas.** Ciência Rural, Santa Maria, v.34, n.1, p.843-850, 2003.

BOHNING, J. J.; MISHRA, T.Q; CHOUDHURY, M.The Raman Effect, **A commemorative booklet by American Chemical Society and the Indian Association for the Cultivation of Science**. Jadavpur, Calcutta, 1998.

BORGES, I. C. **Os desafias da engenharia genética na agricultura: percepções de risco e políticas regulatórias.** Tese (Doutorado em Ciências Econômicas) – Instituto de Economia, UNICAMP, Campinas-SP, 2010.

BORIN, A.; FERRÃO, M.F.; MELLO, C.; CORDI, L.; PATACA, L.C.M.; DURAN,N.; POPPI, R.J. **Quantification of lactobacillius in fermeted milk by multivariate image analysis with least-square spport-vector machines**. Analytical and Bioanalytical Chemistry, v. 387, p. 1105, 2007.

BOUGEAR, D.; SCHADER, B. **Infrared and Raman Spectroscopy : Methods and Applications**. Weinheim: VCH, 1995.

BOVOLINI, M. S. **Análise exploratória de espectros no infravermelho de formulações comerciais contendo glibenclamida compactados por transformadas wavelets.** Universidade de Santa Cruz do Sul. 2010. Dissertação (Mestrado em Sistemas e Processos Industriais) - Universidade de Santa Cruz do Sul, Rio Grande do Sul.

BRAGA, J. W. B.; POPPI, R. J. **Validação de modelos de calibração multivariada: uma aplicação na determinação de pureza polimórfica de carbamazepina por espectroscopia no infravermelho próximo**. Química. Nova, v. 27, n.6, p.1004-1011, 2004.

BRERETON, R. G. **Introduction to multivariate calibration in analytical chemistry**. Analyst, v.125, 2125-2154, 2000.

_____Self organising maps for visualising and modelling. Chemistry Central Journal, v. 6, n, 2012.

_____Chemometrics: Data Driven Extraction for Science. 2. ed. Hoboken, NJ, USA: John Wiley & Sons, 2018.

BRUNS, R. E.; FAIGLES, J. F. G. **Quimiometria.** Química Nova, p. 84-99, 1985.

BURGER, J.; GOWEN, A. **Data handling in hyperspectral image analysis.** Chemometrics and Intelligent Laboratory Systems, v.108, p.13-22, 2011.

BURGER, J.; GELAD, P. **Hyperspectral NIR image regression part I: calibration and correction**. Journal of Chemometrics, v. 19, p. 355–63, 2005.

CANILHA, L.; CARVALHO, W. A. F.; M.D. et al. **Ethanol Production from Sugarcane Bagasse Hydrolysate Using Pichia stipitis.** Appl Biochem Biotechnol, v. 161, p. 84–92, 2010.

CARVALHO, L. P.; ANDRADE, F. P.; SILVA FILHO, J. L. **Cultivares de algodão colorido no Brasil**. Revista Brasileira de Oleaginosas e Fibrosas, v. 15, n. 1, p. 37-44, 2011.

CHEN, Q.; ZHAO, J.; ZHANG, H.; WANG, X. **Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration**. Analytica Chimica Acta, v. 572, p. 77-84, 2006.

CONCEIÇÃO, F.R. et al**. Detecção de organismos geneticamente modificados**. In: BINSFELD, P.C. Biossegurança em biotecnologia. Rio de Janeiro: Interciência. p.145-169, 2004.

COSTA FILHO, P. A. **Estudo comparativo entre técnicas de inteligência artificial e modelos lineares em determinações quantitativas no infravermelho próximo**. 2003. Tese (Doutorado), Universidade Estadual de Campinas- UNICAMP. Campinas, São Paulo.

CRISÓSTOMO, J. R. **Avaliação da estrutura e do potencial genético de uma população de algodoeiro (*G. hirsutum* L.) parcialmente autógama**. Tese (Doutorado) - Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiróz", Piracicaba, 191 f, 1989.

EDWARDS, H. G.; FARWELL, D. W.; WEBSTER, D. **FT Raman microscopy of untreated natural plant fibres**. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, v. 53ª, p. 2383–92, 1997.

ELIERS, P. H. C.; BOELENS, H. M. *F.* **Baseline correction with asymmetric least squares smoothing**. Leiden University Medical Centre Report, 2005.

ESMODE-WHITE, F. W. L; ESMODE-WHITE, K. A.; MORRIS, M. D. **Minor Distortions with Major Consequences: Correcting Distortions in Imaging Spectrographs**. Applied Spectroscopy, v. 65, p. 85–98, 2011.

FAND, B,B.; SUROSHE, S.S. **The invasive mealybug Phenacoccus solenopsis Tinsley, a threat to tropical and subtropical agricultural and horticultural production systems.** A review, v. 69, p. 34-43, 2015.

FARIA, D.L.A.; EDWARDS, H.G.M.; AFONSO, M.C.; BRODY, R.H.; MORAIS, J. L. **Raman Spectroscopic Analysis of a Tembetá: a Resin Archaeological Artefact in Need of Conservation.** Spectrochim. Acta A, submetido, 2002.

FARIA, D. L. A; SANTOS. L.; GONÇALVES, N. **Uma demonstração sobre o espalhamento inelástico de luz: Repetindo o experimento de Raman**. Quim. Nova, 20, 319, 1997.

FAWCETT, T. **An introduction to ROC analysis.** Pattern Recognition Letters, v. 27, n. 8, p. 861–874, jun. 2006.

FERRARO, J. R. **Basic Theory. In: Introductory Raman spectroscopy**. p. 1-95. Ferraro, J. R., Ed., Academic Press, London, 2003.

FERRÃO, M. F. **Aplicação de técnicas espectroscópicas de reflexão no infravermelho no controle de qualidade de farinha de trigo**. 2000. Tese. Universidade Estadual de Campinas – UNICAMP, Campinas, São Paulo.

FERREIRA, M. M. C. **Quimiometria: Conceitos, métodos e Aplicações**. Campinas, SP: Editora da Unicamp, 2015.

FERREIRA, M. M. C. **Quimiometria: Conceitos, métodos e Aplicações**. Campinas, SP: Editora da Unicamp, 2015.

FERRÃO, M. F. **Aplicação de técnicas espectroscopicas de refçexão no infravermelho no controle de qualidade de farinha de trigo**. 2000. Tese. Universidade Estadual de Campinas – UNICAMP. Campinas, São Paulo.

FOWLER, J.L.; RAY, L.L. **Response of two cotton genotypes to five equidistant spacing patterns**. Agronomy Journal, Madison, v.69, n.5, p.733-738, 1977

FRIEDRICHS, M. S. **A model-free algorithm for the removal of baseline artifacts.** Journal of Biomolecular NMR, v. 5, p. 147–153, 1995.

FREDDI, O. S.; STOQUERO, A.; PERES, F.S.C. **Saturação do solo pos base na produtividade e qualidade de fibra do algodoeiro.** Nativa, Sinop, v. 02, n. 01, p. 06-12, 2014.

GAUTAM, R.; VANGA, S.; ARIESE, F.; UMAPATHY, S. **Review of multidimensional data processing approaches for Raman and infrared spectroscopy.** Epj Techniques and Instrumentation 2, 2015.

GELADI, P. KOWALSKI, B. **Partial l a t-square regressio: a tutorial.** Analytical Chimica Acta, v. 185, p. 1, 1986.

GEMPERLINE, P.; Practical Guide to Chemometrics. **CRC Press Taylor & Francis Group**, 2006.

GONZÁLEZ-MARTÍN, M. I.; WELLS MONCADA, G.; GONZÁLEZ-PÉREZ, C.; ZAPATA SAN MARTÍN, N.; LÓPEZ-GONZÁLEZ, F.; LOBOS ORTEGA, I.; HERNÁNDEZ-HIERRO, J. M. **Chilean flour and wheat grain: Tracing their origin using near infrared spectroscopy and chemometrics**. Food Chemistry, v. 145, p. 802–806, 2014.

GONZAGA, A., "**Métodos de avaliação de Classificadores",** 2011.

HAALAND, D. M.; THOMAS, E.V. **Partial Least-Squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information.** Analytical Chemistry, v.60, p. 1193, 1988.

HERRENO, A. M. **Raman Spectroscopy for Monitoring Protein Structure in Muscle Food Systems.** Critical Reviews in Food Science and Nutrition, 48:6, 512-523, 2008.

HOLLER, F. J.; SKOOG, D. A.; CROUCH, S. R. **Princípios de análise instrumental**. 6. ed. Porto Alegre: Bookman, p. 1055, 2009.

HORTON, R. B.; DURANTY, E.; McCONICO, M.; VOGT, F. **Fourier transform infrared spectroscopy and improved principal component regression (PCR) for quantification of solid analytes in microalgae and bacteria.** Appl Spectrosc., v. 65, p. 442–53, 2011.

JABLONSKI, A. **Efficiency of Anti-Stokes Fluorescence in Dyes**. Nature v. 131, p. 839–840, 1933.

JOHSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 3. ed. Englewood Cliffs: Prentice Hall, 1992.

KAMBLE, A. L.; FAND, B. B. **Linking Farmers' Perception of Climate Change to the Changing Agricultural Pest Problems: A Case Study of Cotton Mealybug Phenacoccus solenopsis Tinsley in Maharashtra, India.** International Journal of Current Microbiology and Applied Sciences, v. 8, n. 3, 2019.

KATSUMOTO, Y.; OZAKI, Y. **Practical algorithm for reducing convex spike noises on a spectrum.** Applied Spectroscopy, v.57, p. 317–322, 2003.

KRANTHI, K.R; RUSSEL, D.A. **Changing Trends in Cotton Pest Management.** Integrated Pest Management: Innovation-Development Process, 2009.

LAGO, C. L.; JULIANO, V. F.; KASCHERES, C. **Applying Moving Median Digital Filter To Mass Spectrometry And Potentiometric Titration**. Analytica Chimica Acta., v. 310, n. 2, p. 281 - 288, 1995.

LAU, D. P.; HUANG, Z.; LUI, H.; ANDERSON, D. W.; BEREN, K.; MORRISON, M. D.; SHEN, L.; ZENH, H. **Raman spectroscopy for optical diagnosis in the larynx: Preliminary findings**. Lasers in Surgery and Medicine, v. 37, p. 192–200, 2005.

LAVINE, B. K.; WORKMAN, J. **Chemometrics.** Analytical Chemistry, v. 85, p. 705, 2013.

LEVANDI, T.; LEON, C.; KALJURAND, M.; GARCIA-CAÑAS, V.; CIFUENTES, V. **Capillary electrophoresis time-of-flight mass spectrometry for comparative metabolomics of transgenic versus conventional maize.** *Anal. Chem*., 80:6329-6335,2008.

LI-CHAN, E. C. Y.; NAKAI. S.; HIROTSUKA, M. **Raman spectroscopy as a probe of protein structure in food system.** In: Protein structure. Function relationships in foods. pp.163–197, 1994.

LI, M., et al. **Evaluating growth models of *Pseudomonas* spp. in seasoned prepared chicken stored at different temperatures by the principal component analysis (PCA).** Food Microbiology, v. 40, p. 41–47, 2014.

LIU, J. **Terahertz spectroscopy and chemometrics classification of transgenic corn oil from corn edible oil.** Microwave and optical technology Letters. Vol. 59, Nº 3, March 2016.

LIU, J.; MAO, L.; KU, J.; PENG, H.; LAO, Z.; CHEN, D.; YANG, B. **Using terahertz spectroscopy to identify transgenic cottonseed oil according to physicochemical quality parameters.** Optik, v. 142, p. 483–488, 2017.

LORBER, A. **"Error propagation and figures of merit for quantification by solving matrix equations".** Analytical Chemistry, v. 58, p. 1167-1172, 1986.

LU, B. R.; YANG, C. **Gene flow from genetically modified rice to its wild relatives: assessing potential ecological consequences.** Biotechnology Advances. v. 27, p. 1083–1091, 2009.

LU, Y.; WU, K.; JIANG, Y.; XIA, B.; LI, P.; FENG, H.; WYCKHUYS, K. A.G.; GUO, Y. Mirid bug outbreaks in multiple crops correlated with wide-scale adoption of Bt cotton in China. Ssciense. v. 328, p. 1151–1154, 2010.

LUYPAERT, J.; HEURERDING, S.; HEYDEN, Y.V.; MASSART, D.I. **The effect of preprocessing methods in reducing variability from near-infrared measurements os creams**. Journal of Pharmaceutical and Biomedical Analysis, v. 36, p. 495-503, 2004.

MACHADO, V.A.; POTIN, M. P.; TORRES, J.B. **Selective Insecticides Secure Natural Enemies Action in Cotton Pest Management.** Ecotoxicology and Environmental Safety, vol. 184, p. 109669, 2019.

MALINI, R; VENKATAKRISHMA, K.; KURIEN, J. et al. **Discrimination of normal, inflammatory, premalignant, and malignant oral tissue: A Raman spectroscopy.** study. Biopolymers, v. 81, n. 3, p. 179–193,2006.

MARTENS, H.; NAES, T.; **Multivariate calibration.** Wiley: New York, 1996.

METROHM. **Analisadores Raman portáteis Mira, Misa e SERS.** Disponivel em: https://www.metrohm.com/pt-br/produtos-geral/espectroscopia/espectrometros-portateis-mira/. Acesso em: 10 de jan. 2021.

MEYABADI, T.F.; DADASHIAN, F.; SADEGHI, G.M.; ASI, H.E.Z**. Spherical cellulose nanoparticles preparation from waste cotton using a green method.** Powder Technology, v. 261, 2014.

MORAIS, S. Cristiane. **Aplicação de Calibração Multivariada em dados de Espectroscopia UV-Visível para previsão da Acidez total em vinhos.** 2014. Trabalho de conclusão de curso – Tecnologia em Alimentos, Universidade Tecnológica Federal do Paraná (UTFPR). Campo Mourão, 2014.

MORGANO, M.A. **Aplicação de métodos quimiométricos em análise de alimentos**.2005. Tese (Doutorado em Ciências). Universidade Estadual de Campinas – UNICAMP, Campus, São Paulo.

NAES, T.; MARTENS, H. **Multivariate Callibration.** John Wiley & Sons Ltda, 1989.

NAGATA, N.; BUENO, M. I. M. S.; PERALTA-ZAMORA, P. **G. Métodos matemáticos para correção de interferências espectrais e efeitos interelementos na análise quantitativa por fluorescência de raios-X.** Química. Nova, v. 24, n.4. p. 531-539, 2001.

NEUMANN, N. M.; PLASTIO, A.; PINTO JUNIOR, J. A.; FREITAS, A. **Is p -value 0.05 enough? A study on the statistical evaluation of classifiers**, 2018.

NOTINGHER, I.; GREEN, C.; DYER, C. **Discrimination between ricin and sulphur mustard toxicity in vitro using Raman spectroscopy.** Journal of theRoyal Society Interface, v. 1, p. 79–90, 2004.

O'CONNELL, M. L.; RYDER, A. G.; LEGER, M. N.; HOWLEY, T. **Qualitative analysis using Raman spectroscopy and chemometrics: a comprehensive model system for narcotics analysis.** Appl Spectrosc.,v. 64, p. 1109–21, 2010.

OLIVERI, A.C.; FABER, N. M**. Validation and Error**. In: Comprehensive Chemometrics, 2009; Vol. 3, pp 91−120.

OLIVERI, A.C.; FABER, N. M.; FERRÉ, J.; BOQUÉ. R.; KALIVAS, J. H.; MARK, H. **Uncertainty estimation and figures of merit for multivariate calibration (IUPAC technical report).** Pure Applied Chemistry, v. 78, p. 633−661, 2006.

OLIVERI, P.; DOWNEY, G**. Multivariate class modeling for the verification of foodauthenticityclaims.** TrAC Trends in Analytical Chemistry, v. 35, p. 74–86,2012.

OTTO, M. **Chemometrics: Statistics and Computer Application in Analytical Chemistry.** 2. ed. Alemanha: Wiley-vch, p. 328, 2007.

PENNA, J. C. V. Melhoramento do algodão. In: BORÉM, A. (Ed). **Melhoramento de espécies cultivadas**. 2. ed. Viçosa, MG: UFV, p. 15-53, 2005.

PIZARRO, C.; ESTEBAN-DIEZ, I.; GONZÁLEZ-SAIZ, J. M. **Influence os data pre-processing on the quantitative determination of the ash content lipids in roastes coffe by near infrared spectroscopy.** Analytuca Chimica Acta, v. 509, p. 217-227, 2004.

PETERSON, G.; CUNNINGHAM, S.; DEUTSCH, L.; ERICKSON, J.; QUINLAN, A.; RAEZLUNA, E.; TINCH, R.; TROEL, M.; WOODBURY, P.; ZENS, S. **The risks and benefits of genetically modified crops: a multidisciplinary perspective**. Conservation Ecology, v. 4, n. 1, p. 13, 2012.

PUJER, S,; SIWACH, S.S.; DESHMUKH, J.; SANGWAN, R.S.; SANGWAN, O. **Genetic variability, correlation and path analysis in upland cotton (Gossypium hirsutum L.).** Electronic Journal of Plant Breeding, v. 5, n. 2, p. 284-289, 2014.

QAIM, M., ZILBERMAN, D. **Yield effects of genetically modified crops in Developing Countries.** Science, v. 299, 2003.

RAHMAN. M. H.; WANG, X.; WAJID, A.; NASIM, W.; HUSSAIN, M.; AHMAD, B.; AHMAD I.; ALI, Z.; ISHAQUE, W.; AWAIS, M.; SHELIA, V.; AHMAD, S.; FAHAD, S.; ALAM, M.; ULLAH, H.; HOOGENBOOM, G. **Multi-model projections of future climate and climate change impacts uncertainty assessment for cotton production in Pakistan.** Agricultural Forest Meteorology, v. 253-254, p. 94–113, 2018.

RANDOLPH, T. W. **Scale-based normalization of spectral data**. Cancer Biomark., v.2, p. 135–44, 2006.

RINNAN, A.; BERG, V. D.; ENGELSEN, S. B. **Review of the most common pre-processing techniques for near-infrared spectra.** Trends in Analytical Chemistry, v. 28, n. 10, 2009.

RINNAN, A.; VANDENBERG, F.; ENGELSEN, S. B. **Review of the most common preprocessing techniques for near-infrared spectra,** Trends in Analytical Chemistry, v. 28, n.10, 2009.

ROCHA, P. D. **Metodologia para distinção entre genótipos convencionais e transgênico de algodão utilizando espectroscopia NIR e imagens Hiperspectrais**. 2020. Dissertação (Mestrado em Química). Universidade Estadual da Paraíba- UEPB, Campina Grande-PB.

RYGULA, A.; MAJZNER, K.; MARZEC, K.M; PILARCZYK, M; BARANSKA, M. **Raman spectroscopy of proteins: a review**. Journal of Raman Spectroscopy, v. 44. p. 1061- 1076, 2013.

SACRÉ, P. Y.; De BLEYE, C.; CHAVEZ, P. F.; NETCHACOVITCH, L.; HUBERT, P.; ZIEMONS, E**. Data processing of vibrational chemical imaging for pharmaceutical applications.** Journal of Pharmaceutical and Biomedical Analysis, v.101, p. 123-140, 2014.

SALA, O.; **Fundamentos da Espectroscopia Raman e no Infravermelho**. 2ª ed., Editora Unesp: São Paulo, 2008.

SANTOS, A.; MENEZES, D.; ELLENA, J.; ANDRADE, M. B.**Aplicação da espectroscopia Raman na caracterização de minerais pertencentes a uma geocoleção**. Quim. Nova, v. 42, n. 5, p. 489-496, 2019**.**

SANTOS, W. J. **Manejo das pragas do algodão com destaque para o cerrado brasileiro**. In: FREIRE, E. C. (Ed.) Algodão no cerrado do Brasil. Brasília: ABRAPA, 2007. Cap. 12, p. 403-478.

SATYAMURTHY, P.; JAIN, P.; BALASUBRAMANYA, R.H.; VIGNESHWARAN, N. **Preparation and characterization of cellulose nanowhiskers from cotton fibres by controlled microbial hydrolysis.** Carbohydrate Polymers, v. 83, n.1, p. 232-240, 2014.

SHIPP, D.W.; SINJAB, F.; NOTINGHER, I. **Raman spectroscopy: Techniques and applications in the life sciences.** Adv. Opt. Photon 9, 315, 2017.

SILVA, G. S.; OLIVEIRA, R. A.; QUEIROZ, N.L.; SILVA, M.N.B.; SOUSA, M.F.; SILVA, S.A. **Desempenho agronômico de algodão orgânico e oleaginosas consorciados com palma forrageira.** Revista Brasileira de Engenharia Agricola e Ambiental, v. 17, n. 9, p. 975-981, 2013.

SILVA, M. A. M.; FERREIRA, M. H.; BRAGA, J.W.B.; SENA, M. M. **"Development and analytical validation of a multivariate calibration method for determination of amoxicillin in suspension formulations by near infrared spectroscopy".** Talanta, v. 89, p. 342-351, 2012).

SMITH, G. P. S.; HUFF, G.S.; GORDON, K.C.; **Investigating Crystallinity Using Low Frequency Raman Spectroscopy:** Applications in Pharmaceutical Analysis. Spectroscopy, v. 31, p. 42–50, 2016.

SOUZA, A. M.; POPPI, R. J. **Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte I**. Química. Nova vol.35, n.1, pp.223-229, 2012.

SOUSA, E. P. **Determinação de parâmetros de qualidade de gasolina e estudo de contaminantes utilizando espectros do infravermelho próximo na fase vapor**. 2005. Tese (Doutorado em Ciências), Universidade Estadual de Campinas – UNICAMP, Campinas, São Paulo.

SOUSA, R.C.P.; SMIDERLE, O.J.; SILVA, S.L. **Validação do método de reflectância de infravermelho na determinação do teor de água em grãos de soja**. In: CONGRESSO.

SOVIERZOSKI, M. A. **Avaliação de Descritores Morfológicos na Identificação de Eventos Epileptiformes**. Tese (Doutorado em Engenharia Elétrica) – Programa de Pós-Graduação em Engenharia Elétrica, Florianópolis, 2009.

STEVENS, A.; RAMIREZ-LOPEZ, L. **An introduction to the prospectr package**. [S. l.], 2013. Disponível em: https://cran.r-project.org/web/packages /prospect. Acesso em: 18 out. 2020.

TARIQ, M.; ABDULLAH, K.; AHMAD, S.; ABBAS, G.; HABIB, M.; RAHMAN; KHAN, M. A. **Weed Management in Cotton**. Cotton Production and Uses, 2020.

TAYLOR; FRANCIS. **Pratical guide to chemometrics**. 2. ed. Nova York: CRC, 2006.

ULF, G.; **Near-Infrared Raman Spectroscopy.** Lund Reports on Atomic Physics, 1993.

VALENTINI, S.R.; SOMMER W.A.; MATIOLI, G. **Validação de métodos analíticos**. Arquivos do Mudi, v.11, n. 2, p. 26-31, 2007.

VANDEGINSTE, B. G. M., MASSART, D. L., BUYDENS, L. M. C., DE JONG, S., LEWI, P.J., SMEYERS-VERBEKE, J. **Handbook of Chemometrics and Qualimetrics: Part B.** SCIENCE, E. Amsterdam, 1988.

VESSMAN, J,; STEFAN, R.I.; STADEN, J. F.V.; DANZER, K.; LINDNER, W.; BURNS, D. T.; FAJGELJ, A.; MÜLLER, H. **Selectivity in analytical chemistry**. Pure Applied Chemistry, v. 73, p. 138, 2001.

VIDAL, M.; AMIGO, J. M. **Pre-processing of hyperspectral images. Essential steps before image analysis.** Chemometrics and Intelligent Laboratory Systems, v.117, p.138-148, 2012.
WILLIAMS, R. P.; SWINKELS, A. J.; MAEDER, M. **"Identification and application of a prognostic vector for use in multivariate calibration and prediction".** Chemometrics and Intelligent Laboratory Systems, v. 15, p. 185-193, 1992).

WOLD, S.; SJÖSTRÖM, M; ERIKSSON, L**. PLS-regression: a basic tool of chemometrics.** Chemometrics Intelligent Laboratory Systems, v. 58, p. 109-130, 2001.

XIE, L.; YING, Y.; YING, T.; YOU, H.; FU, X**. Discrimination of transgenic tomatoes based on visible/near-infrared spectra**. Analytica Chimica Acta, v. 584, p. 379–384, 2007.

XU, W.; LIU, L.; XIE, Y.; YNG, Y. **Comparison of Fourier transform near-infrared, visible near-infrared, mid-infrared, and Raman spectroscopy as non-invasive tools for transgenic rice discrimination.** American Society of Agricultural and Biological Engineers. Vol. 57(1), 2014.

XU, W.; XIE, L.; YE, Z.; GAO, W.; YAO, Y,; CHEN, M.; QIN, J.; YING, Y. **Discrimination of Transgenic Rice containing the Cry1Ab Protein using Terahertz Spectroscopy and Chemometrics**. Scientific Reports, 2015.

ZHANG, L.; HENSON, M. J. **A Practical Algorithm to Remove Cosmic Spikes in Raman Imaging Data for Pharmaceutical Applications.** Applied Spectroscopy, v. 61, p. 1015-1020, 2007.

ZHANG, D.; BEN-AMOTZ, D. **"Removal of Cosmic Spikes from Hyper-spectral Images Using a Hybrid Upper-Bound Spectrum Method.** Applied Spectroscopy, v. **56, p. 91-98, 2002.**

ZHAO, J.; LUI, H.; McLEAN, D.; ZHENG, H. **Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy**. Applied Spectroscopy, v.61, p. 1225–32, 2007.

ZHANG, H-B; LI, Y.; WANG, B.; CHEE, P. W. **Recent advances in cotton genomics**. International Journal of Plant Genomics, Cairo, v. 2008, p. 1, 2008. Disponível em:<http://www.hindawi.com/GetArticle.aspx?doi=10.1155/2008/742304>. Acesso em: mar. 2009.

ZOHAIB, A.; TABASSUM, T.; JABBAR, A.; ANJUM, S. A.; ABBAS, T.; MEHMOOD. A.; IRSHAD, S.; KASHIF, M.; NAWAZ, M.; FAROOQ, N.; NASIR, I.

R.; RASOOL, T.; NADEEM, M.; AHMAD, R. **Efect of Plant Density, Boron Nutrition and Growth Regulation on Seed Mass, Emergence and Ofspring Growth Plasticity in Cotton.** Scientific Reports, v.8, n.7953, p. 1-14, 2018.

ANEXO A - CRONOGRAMA DE ATIVIDADES DO TRABALHO DE DISSERTAÇÃO

O cronograma de execução das atividades que fazem parte do desenvolvimento do projeto de dissertação está descrito na tabela 1. As atividades a serem realizadas estão listadas abaixo.

Atividade 1: Levantamento bibliográfico.

Atividade 2: Obtenção das informações sobre a identificação convencional e das espécies geneticamente modificadas.

Atividade 3: Obtenção das espécies convencionais e transgênicas dos cultivares de algodão a serem utilizadas neste estudo.

Atividade 4: Registros dos espectros Raman das sementes convencionais e transgênicas.

Atividade 5: Tratamento dos dados Raman das sementes de algodão, utilizando técnicas de reconhecimento de padrões não-supervisionados.

Atividade 6: Construção dos modelos de classificação utilizando os espectros Raman das sementes de algodão juntamente com técnicas de reconhecimentos de padrões supervisionados.

Atividade 7: Preparação e defesa da qualificação.

Atividade 8: Publicação dos resultados.

Atividade 9: Defesa da dissertação.

Tabela 2 - Cronograma das atividades a serem desenvolvidas durante a vigência (de Mar/2019 a Fev/2021).

VERDE: realizado **VERMELHO:** a desenvolver **LARANJA**: em andamento

PLANEJAMENTO DAS ATIVIDADES	PERÍODO (MESES)											
	01	02	03	04	05	06	07	08	09	10	11	12
Atividade 1	X	X	X	X	X	X	X	X	X	X	X	X
Atividade 2			X	X								
Atividade 3			X	X	X							
Atividade 4			X	X	X							
Atividade 5						X	X	X				
Atividade 6							X	X	X			
Atividade 7								X	X	X		
Atividade 8								X	X	X	X	
Atividade 9												X

ANEXO B – CURRICULUM LATTES

Mayara Macedo da Mata

Endereço para acessar este CV:

http://lattes.cnpq.br/1963709790423217

ID Lattes: 1963709790423217

Última atualização: 18/02/2021

IDENTIFICAÇÃO

Nome: Mayara Macedo da Mata.
Nome em citações bibliográficas: MATA, M. M.
Lattes ID: http://lattes.cnpq.br/1963709790423217

Formação

- Graduação em Licenciatura em Química (2017.1).
  Universidade Estadual da Paraíba, UEPB, Brasil.
- Graduação em andamento em Química Industrial.
  Universidade Estadual da Paraíba, UEPB, Brasil.
- Mestrado em andamento em Química (Conceito CAPES 3).
  Universidade Estadual da Paraíba, UEPB, Brasil. Orientador: Simone da Silva Simões.

Formação Complementar

- A química da produção de Bebidas Fermentadas. (Carga horária: 4h).

  Universidade Estadual da Paraíba, UEPB, Brasil. (2019).

Projeto de Pesquisa

Metodologias para distinção entre genótipos convencionais e geneticamente modificados de algodão utilizando imagens hiperespectrais.

Descrição: Métodos de engenharia genética geralmente são empregados para obter características não conseguidas convencionalmente, como aumento na produtividade, menor uso de herbicidas e adaptação a diversas regiões. Em relação a cultura algodoeira, estas técnicas são utilizadas para resolver alguns problemas como produtividade e resistência a pragas. A identificação de eventos transgênicos pode ser realizada por métodos baseados na sequência do DNA exógeno ou na proteína transgênica, como bioensaios e imunoensaios, que podem ser dispendiosos e demorados. Assim, o presente projeto propõe uma metodologia baseada na espectroscopia RAMAN para a identificação e distinção entre sementes convencionais e geneticamente modificadas (GM) do algodoeiro, de forma rápida e pouco onerosa..

Situação: Em andamento; Natureza: Pesquisa. Alunos envolvidos: Graduação: (3) / Mestrado acadêmico: (2) / Doutorado: (2) .
Idiomas
Português - Compreende Bem, Fala Bem, Lê Bem, Escreve Bem. **Espanhol** - Compreende razoável, Fala razoável, Lê bem, Escreve razoável. **Inglês** - Compreende pouco, Fala pouco, Lê razoavelmente, Escreve pouco.
PRODUÇÕES
Apresentações de Trabalho
• **MATA, M. M.**; SIMOES, S. S. ; MEDEIROS, E. P. ; ROCHA, P. D. ;FARIAS, I. K. T. ; SILVA, J. L. B. ; SILVA, A. G. . ESPECTROSCOPIA RAMAN PARA DISTINÇÃO ENTRE SEMENTES DE ALGODÃO CONVENCIONAIS DE TRANSGÊNICAS. 2019. (Apresentação de Trabalho/Conferência ou palestra).
EVENTOS
Participação em eventos, congressos, exposições e feiras
• I SEMANA DE QUÍMICA DA UEPB.150 ANOS DA TABELA PERIODICA. 2019. (Encontro). • I Workshop Paraibano de Pós- Graduação em Química. Química, Tecnologia e seus Impactos na Sociedade. 2019. (Oficina). • X OLIMPÍADA PARAIBANA DE QUÍMICA. OPBQ & SEMANA DE EDUCAÇÃO EM QUÍMICA. 2019. (Congresso). • VI Encontro de Química da UFPB - Even3. 2019. (Encontro). • II Workshop Paraibano de Pós- Graduação em Química. Química, Tecnologia e seus Impactos na Sociedade. 2020. (Oficina).
Organização de eventos, congressos, exposições e feiras
• SIMOES, S. S. ; LUCENA, J. R. ; ALVES, M. C. F. ; **MATA, M. M.** . I Workshop Paraibano de Pós-Graduação em Química. 2019. (Outro).